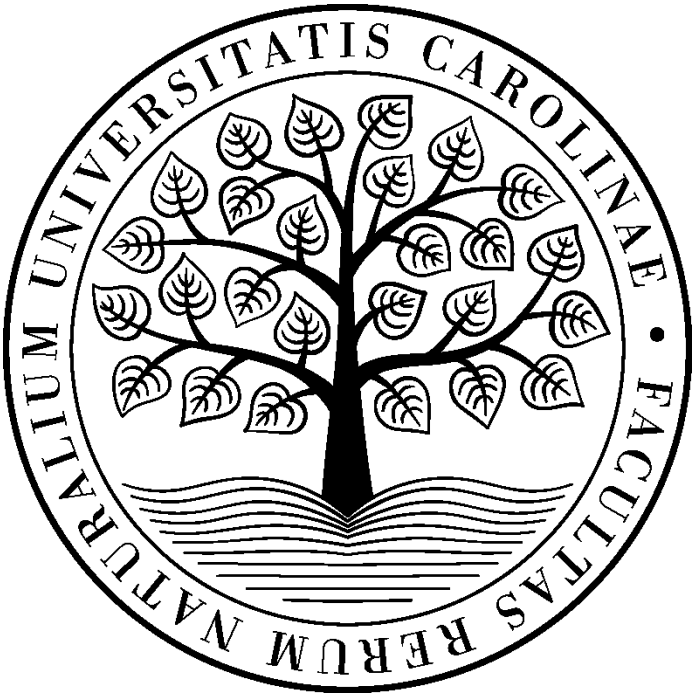
Charles University
Faculty of Science

Study programme: Biology

Branch of study: Theoretical and Evolutionary Biology



Anežka Kielarová

3D struktury fosforylace

3D structures of phosphorylation

Diploma thesis

Supervisor: Mgr. Marian Novotný Ph.D.

Prague, 2019

Acknowledgements

I would like to thank my thesis supervisor Mgr. Marian Novotný, Ph.D. of the Faculty of Science, at Charles University. Without his professional guidance, support and understanding this diploma thesis would have never been accomplished. I would also like to express my very profound gratitude to my husband Petr Kielar for providing me with unfailing support and continuous encouragement throughout my years of study and carrying out research and writing this thesis.

Prohlášení:

Prohlašuji, že jsem závěrečnou práci zpracovala samostatně a že jsem uvedla všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze dne

Podpis: ……

Anežka Kielarová

Abstrakt

Fosforylace je běžná post-translační modifikace proteinů využívaná téměř ve všech buněčných procesech. Přidání fosfátové skupiny na vedlejší řetězec aminokyseliny může z důvodu velikosti fosfátové skupiny a jejího negativního náboje způsobit strukturní změny proteinu a ovlivnit proteinové interakce. Fosforylace také může vést ke změně proteinové funkce, aktivity, a dokonce umístění proteinu v rámci buňky. Experimentální studium fosforylačních míst je velmi časově a finančně náročné i dnes v době hmotnostní spektrometrie. Z tohoto důvodu je předmětem výzkumu mnoha bioinformatických vědeckých skupin predikce fosforylačních míst. Současné analýzy fosforylačních míst studovaly především nefosforylovaná fosforylační místa a rozdělení a zastoupení aminokyselin v jejich sekvenčním okolí. Protože ke specificitě proteinových kináz ale mohou přispívat i aminokyseliny sekvenčně sice vzdálené, ale strukturně blízké, byly v této práci studovány 3D strukturní vlastnosti fosforylačních míst. Zároveň byla poprvé rozsáhle zkoumána fosforylační místa ve fosforylovaném stavu a výsledky byly srovnány s fosforylačními místy v nefosforylovaném stavu. Fosforylační místa byla nalezena především ve smyčkách a na povrchu proteinů. Aminokyseliny v jejich okolí byly častěji hydrofilní, pozitivně nabité a méně blízko sebe než aminokyseliny v okolí nefosforylačních míst. Zároveň bylo zjištěno, že fosforylační místa nejsou signifikantně více konzervovaná v evoluci než nefosforylovaná místa a fosforylace fosforylačních míst způsobuje v 37 % proteinových struktur velké konformační změny. Informace o strukturním okolí fosforylovaných fosforylačních míst by mohla být využita ke zlepšení predikčních nástrojů fosforylačních míst.

Klíčová slova: fosforylace, strukturní charakterizace fosforylačních míst, predikce fosforylačních míst, konformační změny proteinových struktur

Abstract

Protein phosphorylation is a common post-translational protein modification used in almost all cellular processes. When a phosphate group is added to an amino acid side chain, it may alter the protein conformation and protein-protein interactions due to its size and its negative charge. It may also change the protein function, activity and even localization within the cell. Experimental detection of phosphorylation is still extremely labor demanding and very expensive, even when deploying protein mass spectrometry. For this very reason many bioinformatics scientific groups focus on the prediction of protein phosphorylation sites. Recent analyses of phosphorylation sites studied mainly non-phosphorylated phosphorylation sites and the distribution and representation of amino acids sequentially neighboring them. Since sequentially more distant, but structurally close amino acids can contribute to the recognition of protein substrate by protein kinase, structural environment of phosphorylation sites was studied in this thesis. Furthermore, 3D structures of phosphorylation sites were comprehensively studied for the first time in a phosphorylated state and the results were compared with the results obtained from the analysis of non-phosphorylated sites. Phosphorylation sites were found mostly within loops and on protein surfaces. Amino acids spatially neighboring phosphosites were more frequently hydrophilic, positively charged and less close to each other than those neighboring non-phosphorylated residues (serine, threonine and tyrosine). Also, phosphorylation sites were found not to be significantly more conserved in evolution than non-phosphorylated residues. It was also found out that in about 37 % of the cases phosphorylation caused large conformational changes in protein structures. Information on the 3D structural neighborhood of phosphorylated phosphorylation sites can improve the protein phosphosites prediction tools.

Key words: phosphorylation, structural characterization of protein phosphorylation site, protein phosphorylation sites prediction, conformational changes of protein structures

List of abbreviations

aa = amino acid

ANNs = artificial neural networks

ANOVA = Analysis of variance

ATP = adenosine triphosphate

BDT = Bayesian decision theory

CC = correlation coefficient

CRFs = conditional random fields

df = degree of freedom

F1 = F1-score

GTP = guanosine triphosphate

H_0 = zero hypothesis

HMM = hidden Markov model

nonP = non-phosphorylated residues (threonine, serine, tyrosine)

phosphosite = phosphorylation site

pP = phosphorylated phosphosite

Pre = precision

pSer = phosphorylated serine

PSSM = position-specific scoring matrices

pThr = phosphorylated threonine

PTM = posttranslational modification

pTyr = phosphorylated tyrosine

RF = random forest

RMSD = root-mean-square deviation of atomic positions

Ser = serine

Sn = sensitivity

Sp = specificity

SVM = support vector machine

Thr = threonine

Tukey's (HSD) = Tukey's honestly significant difference

Tyr = tyrosine

wP = non-phosphorylated phosphosite

3D = three-dimensional

Table of contents

1. Introduction

Protein phosphorylation is a ubiquitous mechanism of post-translational protein modification. This mechanism is used for protein function regulation through changes of protein conformation, interactions and localization. Being one of the most important reversible mechanisms of post-translational modifications, accurate annotation of sites where phosphorylation can occur (phosphosites) is an important part of cell signaling pathways studies, especially in Eukaryotes (Cohen, 2002). Considering that protein phosphorylation alterations are often associated with global human diseases (e.g. chronic myelogenous leukemia, acute lymphocytic leukemia, gastrointestinal stromal tumors or systemic mastocytosis), phosphosites annotation is also of fundamental importance to biomedical biology (Paul and Mukhopadhyay, 2004).

A high number of phosphosites (e.g. 230 000 within human proteome (Vlastaridis et al., 2017) make experimental identification of an individual phosphosite difficult and time consuming. Given the high number of potential phosphosites, efforts to experimentally identify and verify them all remain challenging. Therefore, development of computational methods to predict potential phosphosites *in silico* is nowadays in the spotlight of many bioinformatics research groups (Hjerrild and Gammeltoft, 2006).

A wide range of algorithms have been used to implement prediction strategies. The main bottleneck of the pattern recognition methods has been the fact that they are often based on very limited experimental data. Moreover, present methods often rely on the properties of a local protein sequence such as the proportion and order of neighboring amino acids. However, the significantly increasing number of experimentally determined phosphosites together with the available three-dimensional (3D) structures of the associated proteins now allow detailed study of protein phosphosites properties in 3D structural context. Information about 3D environments of phosphosites may improve current phosphosites prediction methods (Durek et al., 2009).

The aim of this thesis is to compare three-dimensional structural (context) information of phosphorylated phosphosites with non-phosphorylated phosphosites and to sites where phosphorylation is not known to occur in order to answer the question whether and which structural data can be used, and which properties could be useful for improving protein phosphorylation prediction tools.

2. Literature overview

2.1 Phosphorylation

Post-translational protein modifications (PTMs) can be permanent (e.g. proteolytic cleavage events) or transient (e.g. phosphorylation, glycosylation). Protein phosphorylation as a reversible covalent post-translational modification is performed by the protein kinase and may be reversed by the action of a complementary acting enzyme, phosphatase. The term 'phosphorylation' describes a reaction when a phosphate moiety is transferred from adenosine/guanosine triphosphate (ATP/GTP) to the acceptor residue - a polar side chain of amino acid (Fig. 1). Although virtually any of the amino acids (aa's) with a polar group may be modified, there are mainly three acceptor aa's in Eukaryotes: serine (Ser), threonine (Thr) and tyrosine (Tyr) (Blom et al., 2004).

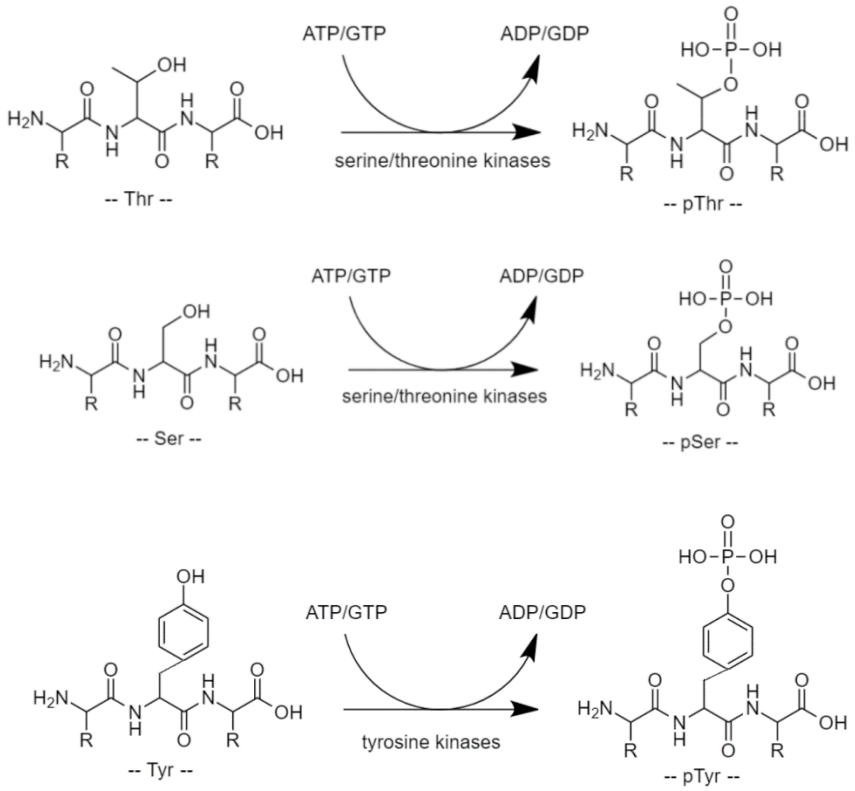


Figure 1: The scheme of protein phosphorylation reactions. The phosphorylation of threonine (top) and serine (middle) is catalyzed by serine/threonine kinases, of tyrosine (bottom) by tyrosine kinases. Drew using ChemDraw (Cousins, 2005).

Post-translational modifications work in complicated and concerted manners. This can be shown on the phenomenon of 'multiphosphorylation'. Cesaro and Pinna in 2015 analyzed phosphosites from PhosphoSitePlus database (Hornbeck et al., 2015) searching for phosphosites that were present on the same protein in a row, forming clusters of three or more phosphorylated residues. They found 631 triplets, 199 quadruplets, 27 quintuplets of phosphorylated serine residues (pSer) and 22 rows composed of more than 5 consecutive pSer. These negatively charged clusters are important for determining and/or

tuning the interactions with other macromolecules (either proteins or nucleic acids) and/or to bind cations and small molecules. This tendency of pSer to cluster close to each other together with hierarchical phosphorylation (once phosphorylated these serine residues will start a downstream cascade of phosphorylation events) highlighted the complexity of phosphorylation as a tool for signaling cascades (Cesaro and Pinna, 2015).

Several mechanisms facilitate phosphorylation event specificity, e.g. the structure of protein kinases active-site clefts, the charge and hydrophobicity of related aa's, multisite phosphorylation of substrate, competition of substrates, or localization of protein kinases and phosphatases. Protein kinases can even serve as scaffold proteins for other protein kinases and phosphatases and this recruiting of phosphorylation components at one moment in one place can lead to ultrasensitive responses or pulses of activity (Roach, 1991)(Ubersax and Ferrell, 2007). Because protein phosphorylation can alter and be altered by another PTMs (so called 'crosstalk' between PTMs), cells can precisely respond to environment stimuli (Duan and Walther, 2015).

Protein kinases are common signaling molecules and protein phosphorylation is often a component of the cell signaling pathways. Manning et al. in 2002 identified 518 protein kinase genes and 106 protein kinase pseudogenes in the human genome. Protein kinases were suggested to constitute 1.7 % of all genes in the human genome (Manning et al., 2002a). There were 10 456 phosphoproteins containing 86 181 phosphosites found in human proteome and 6 512 phosphoproteins containing 36 438 phosphosites were found in mouse proteome. However, the current estimates put the number of phosphorylation events even higher – there were 13 000 phosphoproteins containing 230 000 phosphosites estimated in human proteome, 11 000 phosphoproteins containing 156 000 phosphosites estimated in mouse proteome, yeast (*S. cerevisiae*) were supposed to have 3 000 phosphoproteins and 40 000 phosphosites (Vlastaridis et al., 2017). PhosphoSitePlus (Hornbeck et al., 2015) contained on 17. 4. 2019 173 854 pSer, 71 658 phosphorylated threonine residues (pThr), and 44 633 nonredundant phosphorylated tyrosine residues (pTyr) within 20 443 proteins.

Protein kinases are sorted according to the nature of their substrate on tyrosine kinases, tyrosine-kinase like proteins and serine/threonine kinases. Minor residues are phosphorylated as reaction intermediates or are the products of auto-phosphorylation not catalyzed by kinases, e.g. alkaline phosphatase, phosphoglucomutase, and phosphomannomutase (Strumillo et al., 2018). Most protein kinases families are conserved throughout metazoans (Manning et al., 2002b). A 'typical' protein kinase domain is bilobal with a smaller N-terminal lobe and a bigger C-terminal lobe. While C-terminal lobe is mostly α helical, N-terminal lobe consists of five antiparallel beta sheets (β1-5) and one conserved α-helix. The N-terminal lobe contains an active-site cleft, where the ATP binding site is located. The protein substrate is positioned near this cleft (Fig. 2). Catalysis is mediated by opening and closing of this active-site cleft (De Oliveira et al., 2016).

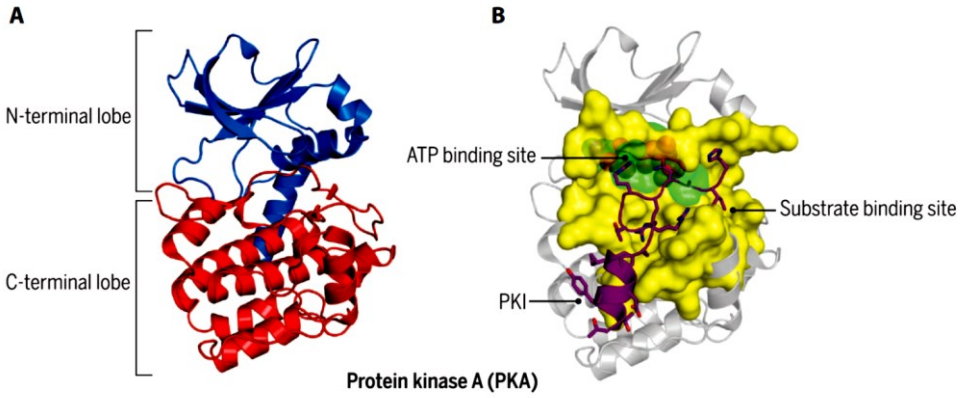


Figure 2: The scheme of 'typical' protein kinase structure illustrated on protein kinase A (PKA). (A) Protein kinase structure consists of two lobes, N- (blue) and C-terminal (red). (B) Between these lobes is an active-site cleft including an ATP binding site (green) and substrate binding sites. The substrate is mimicked here by a protein kinase inhibitor (purple). Adopted from De Oliveira et al. 'Revisiting protein kinase–substrate interactions: Toward therapeutic development' (De Oliveira et al., 2016).

2.2 Phosphorylation sites structural properties

2.2.1 Properties of phosphorylated amino acids

Protein phosphorylation in Eukaryotes is mainly associated with serine, threonine and tyrosine residues. Each residue has its specific characteristics. Threonine has in comparison to serine the additional methyl group on Cβ that controls the side-chain flexibility and can infer with the intramolecular hydrogen bond formation (Kim et al., 2011). Tyrosine is larger than serine and threonine and its aromatic ring mediates more hydrogen bonds and π interactions than serine and threonine side chains. Because phosphate on tyrosine is linked to the O^4 position of the phenolic ring, it lies farther away from the peptide backbone than the phosphate on the β-OH groups of serine and threonine (Fig. 3). Proteins interacting with this pTyr can have deep active-site cleft that can contribute to the specificity of the substrate recognition by tyrosine protein kinases. However, the specificity imparted by the active-site cleft is not absolute. As reviewed in 'Mechanisms of specificity in protein phosphorylation', whereas several serine and threonine protein kinases can phosphorylate tyrosine residues, only few examples were found of the Ser and Thr residues phosphorylated by tyrosine protein kinases (Ubersax and Ferrell, 2007).

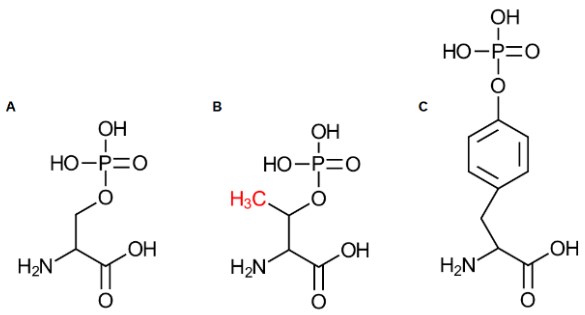


Figure 3: Molecular structures of phosphorylated serine (A), threonine (B) and tyrosine (C). Threonine differs from serine in a methyl group (red) attached on Cβ. Drew using ChemDraw (Cousins, 2005)

2.2.2 Properties of phosphorylation sites 3D environment

Protein kinase recognizes not only a short sequence of the substrate, but also the local 3D environment of phosphorylation sites (phosphosites). Residues localized far from a phosphosite may also contribute to the specificity of the kinase-target interaction. For example, aa's bearing a positive charge define the electrostatic potential surface needed for binding the protein kinase while these positively charged aa's may be localized sequentially distant from the phosphosites, but spatially close (Durek et al., 2009). This can be demonstrated on the phosphorylated tyrosine at position 77 (pTyr77) of *Deinococcus radiodurans* recombinase A protein that interacts with proline 218, arginine 234, aspartate 236, and valine 260 (Fig. 4) (Rajpurohit et al., 2016).

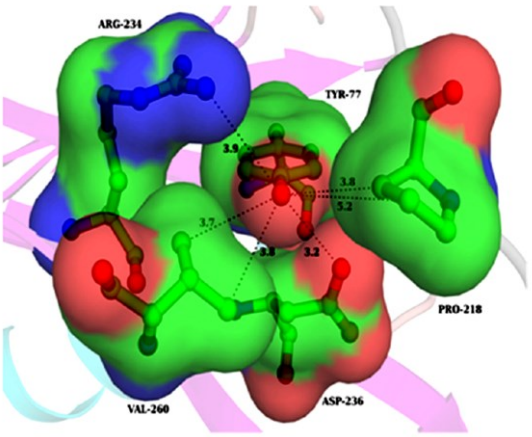


Figure 4: The phosphorylation site of *Deinococcus radiodurans* recombinase A protein (Protein Data Bank entry 1XP8), an example of three-dimensional properties of a phosphorylation site. Phosphorylated tyrosine (pTyr77) interacts with proline 218, arginine 234, aspartate 236, and valine 260 that are sequentially distant, but spatially close. Adopted from Rajpurohit et al. 'Phosphorylation of *Deinococcus radiodurans* RecA Regulates Its Activity and May Contribute to Radioresistance' (Rajpurohit et al., 2016).

Iakoucheva et al. presented the results of protein phosphorylation sites characterization in 2004. They analyzed 613 pSer, 140 pThr, and 136 pTyr and found that the aa's surrounding all three types of phosphorylation sites were mainly exposed to a solvent (localized on the protein surface, not buried within the protein structure). In addition, these aa's had highly flexible side chains (high B-factor) and were mostly hydrophilic and negatively or positively charged (phosphosites were depleted in the uncharged (neutral) residues) (Fig. 5). They also found that phosphosites were more often sequentially surrounded with aa's that were known to be a disorder-promoting (arginine, lysine, glutamic acid, proline and serine) and less with order-promoting aa's (cysteine, tryptophan, tyrosine, isoleucine and valine). So, phosphosites were hypothesized to be preferably present within disorder regions and on the protein surfaces (Iakoucheva et al., 2004).

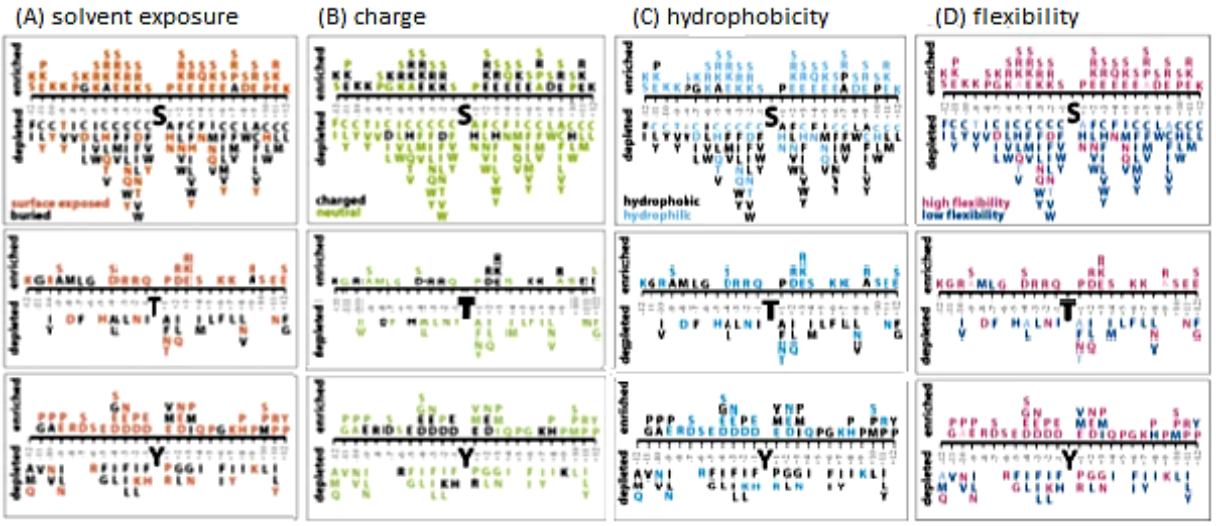


Figure 5: A representation of amino acids spatially neighboring phosphosites. A phosphorylated residue (pSer on top, pThr in the middle, and pTyr on bottom) is located in the middle of the window. Enriched amino acids are above the axis, depleted below it. The representation is shown in each column for the surface exposure (A), charge (B), hydrophobicity (C), and flexibility (D). Adapted from Iakoucheva et al. 'The importance of intrinsic disorder for protein phosphorylation' (Iakoucheva et al., 2004).

In 2005, Fan and Zhang studied 80 biochemical properties of protein phosphorylation site microenvironments to observe features typical for phosphosites. They collected 42 pSer, 19 pThr, and 39 pTyr in resolved protein structures. They used Ser, Thr and Tyr from the same protein structures that were not annotated as phosphorylated as negative samples. After the extraction of features they did 1 000 permutations to test whether these properties appeared randomly or not. They found depletion of isoleucine, phenylalanine, positively charged residues, and residues including an aromatic ring around pSer as well as a reduction of the mobility of aa's spatially surrounding pSer. Around pThr they found the lack of valine residue that could be correlated with the general depletion of hydrophobic and nonpolar residues. Besides, aa's around pThr were more compactly localized (aa's were closer to each other than around non-phosphorylated Thr sites). Tyrosine phosphosites were enriched in positively and negatively charged aa's and notably deficient in cysteine and proline aa's (Fan and Zhang, 2005).

Gnad et al. analyzed a solvent accessibility as well as the tendency of phosphosites to be located in certain protein secondary structures. They collected 1 044 human phosphoproteins and a random set of 998 human proteins. Phosphosites were then annotated for both solvent accessibility and protein secondary structure by SABLE 2.0 program (Adamczak et al., 2005). The results of the analysis showed that phosphosites tend to be solvent accessible, and they occurred predominantly in hinges and loops (Fig. 6) (Gnad et al., 2007).

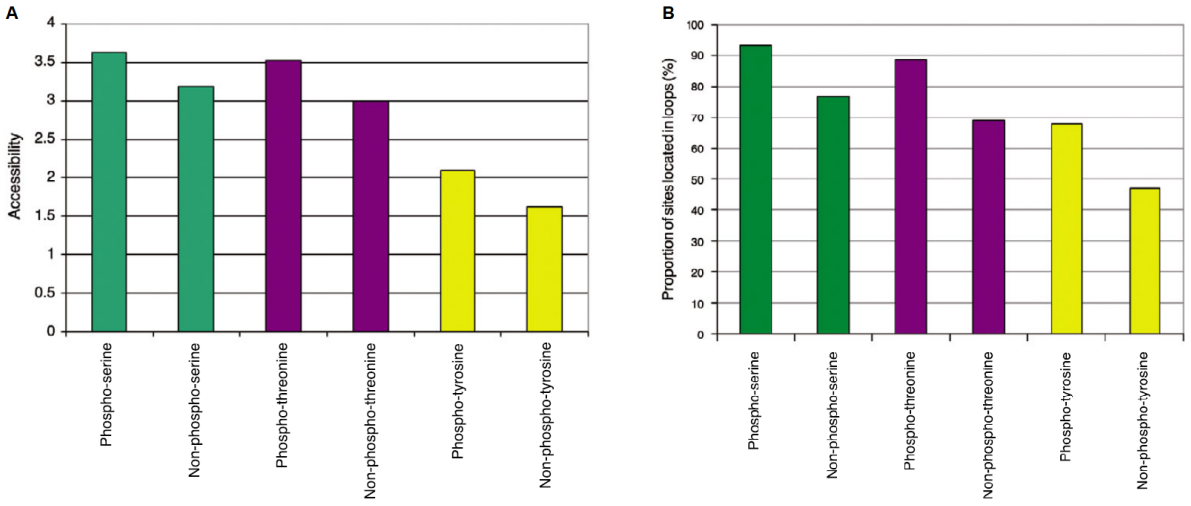


Figure 6: (A) The solvent accessibility of phosphosites (pSer, pThr and pTyr) and non-phosphorylated residues as calculated by SABLE 2.0 (Adamczak et al., 2005). The output of SABLE program was a value on the scale from 0 to 9 (in the range of where 0 means a fully buried and 9 a fully exposed residue. Phosphosites were significantly more solvent exposed than non-phosphorylated sites. (B) The proportion of phosphosites (pSer, pThr and pTyr) and non-phosphorylated sites localized within hinges and loops determined by SABLE 2.0 (Adamczak et al., 2005) and expressed as a percentage. Phosphosites were more frequently located within hinges and loops. Adapted from 'PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites' (Gnad et al., 2007).

In the same year, Jiménez et al. studied structural properties of 264 pSer/pThr and 219 pTyr in 324 nonredundant structural models from MitoCheck's post-translational modifications (mtcPTM) database which had collated human and mouse phosphoproteins. In an agreement with the previous findings of Iakoucheva et al. (Iakoucheva et al., 2004), pSer and pThr were localized mainly in flexible loops and linkers between domains and only marginally (10 %) in rigid secondary structural elements such as α-helices and β-sheets. No tendency to occur more frequently in loops was detected for pTyr. Furthermore, a small number of phosphosites (15 % of studied phosphosites) was buried within the protein structure and thus was not exposed to the solvent (Jiménez et al., 2007).

These results were replicated by Durek et al. in 2009 using 750 non-redundant, structurally resolved phosphosites (363 pSer, 134 pThr, and 253 pTyr). They analyzed structural properties such as secondary structural assignments, relative side chain solvent accessibility and the crystallographic B-factor as a measure of local structural rigidity of phosphorylated and non-phosphosites. A tendency to be more exposed to solvent prooved significant for pSer and pTyr. PThr showed this tendency too, but not significantly. Phosphosites were more often found associated with the largest crystallographic B-Factor, confirming that phosphorylation occurs mainly in regions of greater structural flexibility, albeit significant differences were observed only for pSer (Durek et al., 2009).

The tendency of phosphosites and surrounding amino acids to be located within disordered regions of proteins were confirmed one year later by Gao et al. They collected 61 448 pSer, 14 478 pThr, and 5 727 pTyr across six model organisms (*Homo sapiens, Mus musculus, Drosophila melanogaster, Caenorhabditis elegans, Saccharomyces cerevisiae, and Arabidopsis thaliana*). They analyzed the

output from a protein disorder predictor as well as the frequency of a particular amino acid sequentially surrounding the phosphosite. Their results were in agreement with the previously published results (Iakoucheva et al., 2004). While proline, arginine, aspartate, glutamate, serine, lysine, and glycine were enriched in the surrounding of pSer and pThr, cysteine, tryptophan, tyrosine, phenylalanine, isoleucine, methionine, leucine, histidine, threonine, and valine were depleted. Aspartate, glutamate, proline, serine, and glycine were enriched in vicinity of pTyr, whereas tryptophan, cysteine, phenylalanine, leucine, histidine, methionine, and isoleucine were depleted (Fig. 7) (Gao et al., 2010).

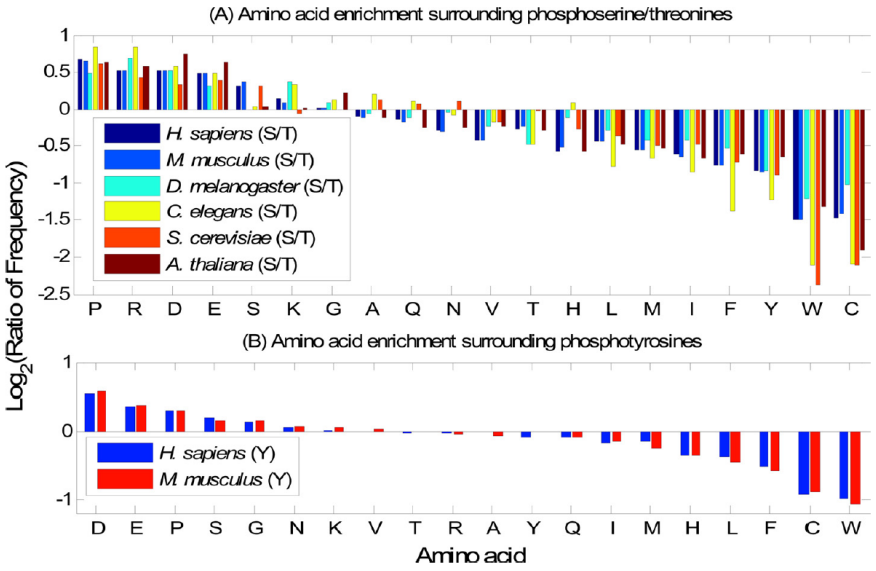


Figure 7: Frequencies of amino acids surrounding phosphosites (from position -6 to +6, where 0 was a phosphorylated residue). On the horizontal axis are one letter codes of amino acids, on the vertical axis the ratio of frequency on a logarithmic scale. (A) Frequencies of amino acids surrounding phosphosites in six model organisms (*Homo sapiens, Mus musculus, Drosophila melanogaster, Caenorhabditis elegans, Saccharomyces cerevisiae, and Arabidopsis thaliana*). (B) Frequencies of amino acids in two well studied model organisms - *Homo sapiens* and *Mus musculus* (Gao et al., 2010).

Palmeri et al. conducted an analysis of *Leishmania* phosphoproteomics data (966 pSer and 210 pThr) and found that phosphosites were more frequently localized in predicted loops (83 %) compared with non-phosphorylated sites (66 %). Besides, 44 % of phosphosites were located within predicted disorder regions compared with 30 % of the non-phosphorylated residues (Zilberstein et al., 2011)). These results were in agreement with (Iakoucheva et al., 2004) and (Jiménez et al., 2007).

Tyanova et al. used a different approach to prove that phosphorylation sites occur preferentially in disordered regions. They analyzed over 5 000 phosphosites (of all three types) on human HeLa S3 cells during the cell cycle (in phases G1, G1S, Early S, Late S, G2 and M) and studied whether phosphorylated residues were located in disordered or ordered regions of proteins and where phosphorylation occurred more frequently. They found that over 90 % of the phosphosites (4 675 sites from 5 173; especially multiple phosphorylation sites) occurred in disordered regions. Furthermore, phosphosites within disordered regions were more often an object of phosphorylation than phosphosites located in ordered regions (Tyanova et al., 2013).

Frades et al. in 2015 showed again the tendency of phosphosites to have less hydrophobic residues in the sequence neighborhood. They identified sequence motifs for phosphoproteomics datasets of *Toxoplasma gondii, Plasmodium falciparum, Schizo-saccharomyces pombe, Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Mus musculus*, *Homo sapiens, Oryza sativa,* and *Arabidopsis thaliana*. Even though the distribution of phosphosites neighboring hydrophobic residues differed slightly between kingdoms/phyla, the phosphosites were in general surrounded with hydrophilic residues and were exposed to solvent (Frades et al., 2015).

In 2015, Huang et al. presented properties of aa's sequentially surrounding the phosphosites across viruses. They collected 2 444 pSer, 635 pThr, and 268 pTyr. They used position weight amino acids composition (PWAA) and found that proline was enriched around phosphosites, especially one position from pSer/pThr from the C-terminal end of the protein. Arginine was enriched in N-term positions of pSer. On the contrary, aspartate and glutamate were enriched in C-term residues of pSer. Moreover, aspartate and glutamate were the dominant amino acids in the flanking of pTyr sites. The analysis of the phosphosites showed that hydrophobic amino acids were present around pSer (36.1 %), which was less than around non-phosphorylated serine (42.9 %). On the contrary, the percentage of acidic amino acids for pSer was 5.7% higher than that for non-phosphorylated serine. They found no significant difference in the representation of acidity, hydrophobicity or polarity between pThr and threonine sites. Because of the small amount of data for pTyr residues, an analysis of pTyr was not conducted (Fig. 8) (Huang et al., 2015).

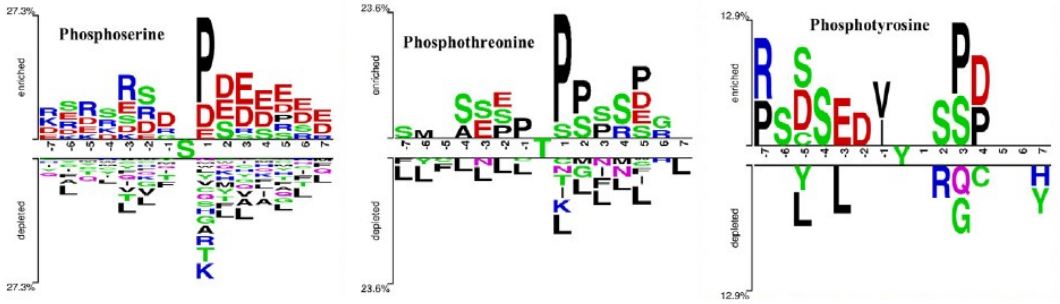


Figure 8: The representation of amino acids around phosphorylated serine (on the left), phosphorylated threonine (in the middle), and phosphorylated tyrosine (on the right). The position of phosphosites is 0, in the middle of the window. Only amino acid residues significantly enriched or depleted (P-value <0.05; t-test) 7 positions up and down from phosphorylation sites were shown. Enriched residues are above the axis, depleted are below it. Adapted from Huang et al. 'Using support vector machines to identify protein phosphorylation sites in viruses' (Huang et al., 2015).

Recently, Karabulut & Frishman analyzed sequence as well as structural properties of 423 pSer, 140 pThr, and 46 pTyr. They found that proline was enriched at position +1 from pSer and pThr. Glutamic and aspartic acids, and serines were also enriched in the upstream regions of pSer, pThr, and pTyr. In the sequence neighborhood of pSer and pThr lysine and arginine were enriched except for the positions from +1 to +4 from the phosphosite. Furthermore, phosphosites with the residues surrounding them up to ten positions from the N- and C- term (the window consists of 21 aa's, phosphosite in the

middle) were significantly more solvent exposed than non-phosphorylated residues. Last but not least, pSer and pThr displayed a tendency to reside in loops (Karabulut and Frishman, 2016).

2.3 Conservation of phosphorylation sites in evolution

Phosphorylation depends not only on properties described above and further analyzed in this thesis, but also on the cell environment. For example, different particles such as scaffold proteins, protein kinases, their substrates and ligands have to be recruited to the same place simultaneously. Because the cellular context plays a fundamental role in the determination of the substrate specificity, more holistic approach is needed (Palmeri et al., 2014). In this section several papers focusing on the incorporation of evolutionary information into the phosphorylation prediction tools will be introduced.

Mann et al. studied the phosphoproteome of *E. coli* and compared its evolutionary conservation with the phosphoproteomes of Eukaryotes, Bacteria and Archea. They used 9 archeal, 53 bacterial and 8 eukaryotic species, and found similarities between the phosphoproteomes of *E. coli* and *B. subtilis* in size, distribution of pSer, pThr, and pTyr, and classes of phosphoproteins. Furthermore, proteins from *E. coli* and *B. subtilis* including phosphorylated residue(s) were generally more conserved than non-phosphorylated proteins in Eukaryotes, Archaea and Bacteria (Fig. 9) (Mann et al., 2007).

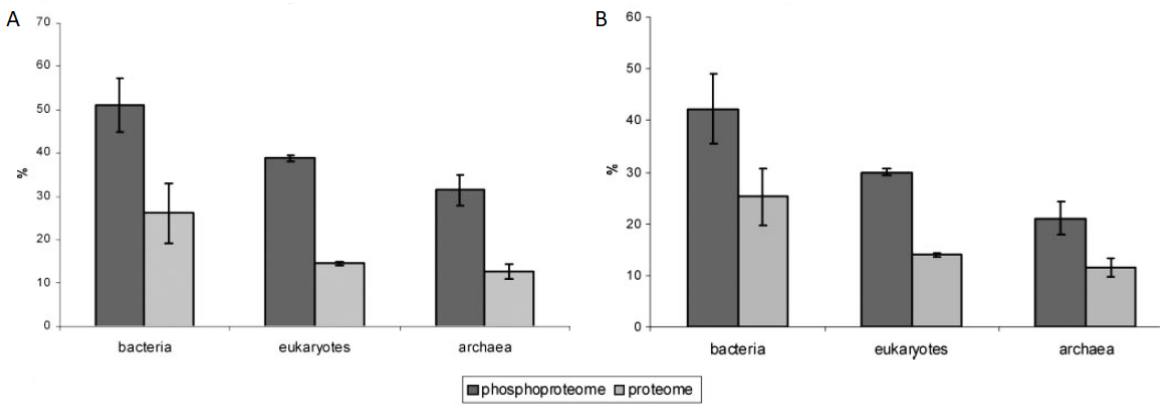


Figure 9: The difference between the evolutionary conservation of the phosphoproteome and the the proteome of (A) *E. Coli*, and (B) *B. Subtilis*. Datasets of the phosphoproteomes and proteomes of 9 archaeal, 53 bacterial, and 8 eukaryotic species were used to perform the evolutionary analysis. Phosphoproteomes showed generally higher evolutionary conservation compared to proteomes.

Gnad et al. analyzed the rate of the conservation of phosphosites and non-phosphorylated sites that occurred within hinges and loops. They analyzed the phosphoproteomes of yeast (*S. cerevisiae*), fly (*D. melanogaster*), zebrafish (*D. renio*), chicken (*G. gallius*), protist (*B. bovis*), rat (*R. norvegicus*), and mouse (*M. musculus*). They found that the conservation of phosphosites is higher than that of non-phosphorylated sites for all three types of phosphorylated residues (pSer, pThr, and pTyr). Nevertheless, threonine is generally less conserved than serine and the higher conservation of pTyr was not significant due to their low number (Fig. 10) (Gnad et al., 2007).

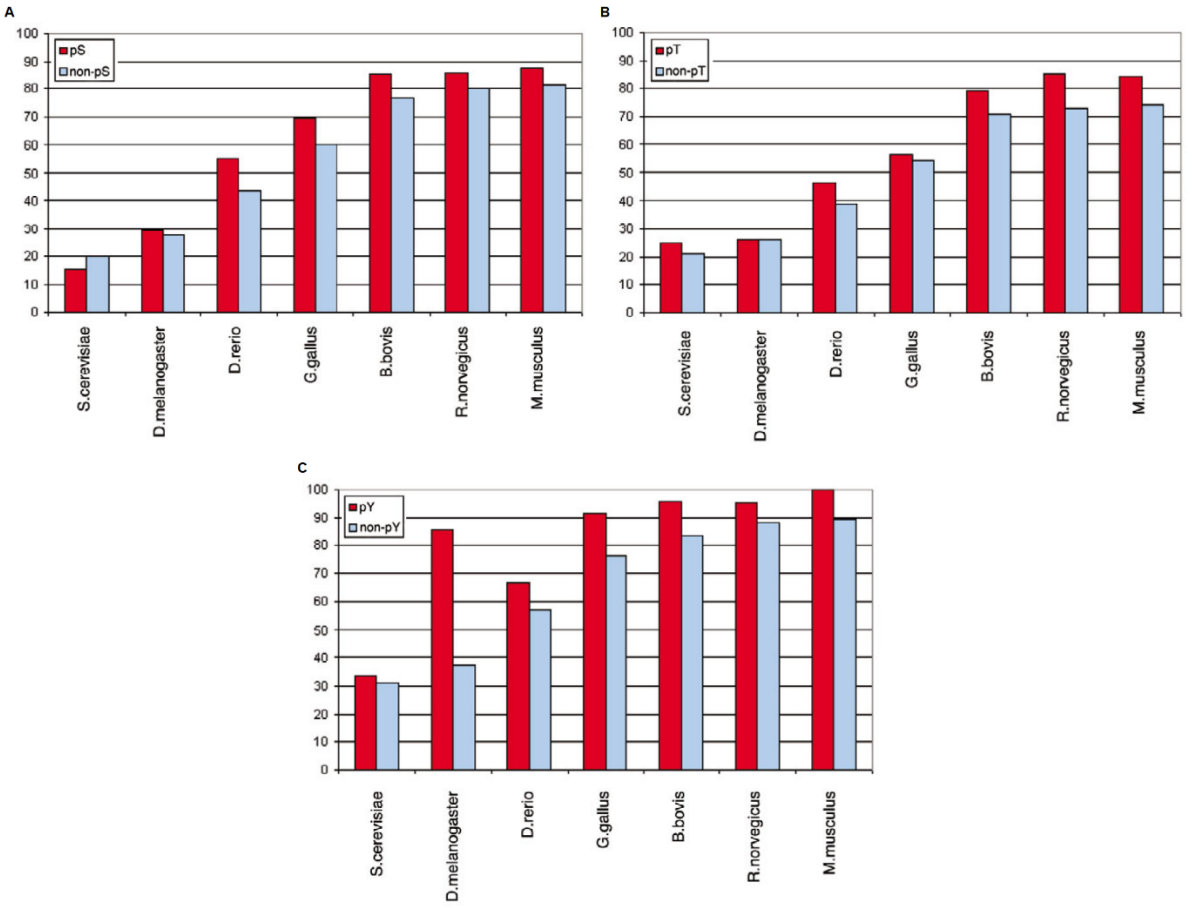


Figure 10: The comparison of the conservation between phosphosites (pSer, pThr and pTyr, red) and non-phosphorylated sites (blue) for yeast (*S. cerevisiae*), fly (*D. melanogaster*), zebrafish (*D. renio*), chicken (*G. gallius*), protist (*B. bovis*), rat (*R. norvegicus*), and mouse (*M. musculus*) expressed as a percentage. (A) Phosphorylated serine residues (pSer) compared to non-phosphorylated serine residues. In all vertebrates, pSer were significantly more conserved than serine. (B) Phosphorylated threonine residues (pThr) compared to non-phosphorylated threonine residues. PThr were significantly more conserved within mammals. (C) Phosphorylated tyrosine residues (pTyr) compared to non-phosphorylated tyrosine residues. Higher conservation level of pTyr was not significant due to their low number. Adapted from 'PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites' (Gnad et al., 2007).

In 2008, Pincus et al. suggested that signaling pathways based on tyrosine phosphorylation evolved shortly before the divergence of metazoans and choanoflagellates and before the evolution of metazoan multicellularity. Therefore, choanoflagellates evolved tyrosine phosphorylation signaling pathways differently from metazoans and so choanoflagellates now contain lineage-specific domain combinations. Moreover, pTyr signaling is in choanoflagellates used for different sets of functions (Pincus et al., 2008).

Tan et. al presented in 2009 that the selection of unintended phosphorylation events that are harmful to the cell may lead to progressive depletion of phosphosites during evolution. They analyzed the genomes of distinct species: yeast (*S. cerevisiae*), worm (*C. elegans*), sea squirt (*C. intestinalis*), fly (*D. melanogaster*), mosquito (*A. gambiae*), zebrafish (*D. rerio*), tetraodon pufferfish (*T. nigroviridis*), Japanese pufferfish (*T. rubripes*), frog (*X. tropicalis*), mouse (*M. musculus*), rat (*R. norvegicus*), chicken (*G. gallus*), dog (*C. familiaris*), cow (*B. taurus*), chimpanzee (*P. troglodytes*) and human (*H. sapiens*).

They showed that Metazoans lost a great amount of tyrosine residues after the invention of the signaling mechanisms based on tyrosine phosphorylation (Fig. 11). Because the number of tyrosine kinases predicted from genomic data increased in evolution of multicellularity, they suggested that it can be a result of the effort to minimize the noise in signaling networks and thus eliminate unspecific detrimental phosphorylation events by tyrosine-removing mutations and the evolution of more specific tyrosine kinases (Tan et al., 2009). Since this selection pressure may have existed, Creixell et al. hypothesized that not all unintended phosphorylation events are necessarily damaging the cell (otherwise they would be depleted in evolution) (Creixell et al., 2012).

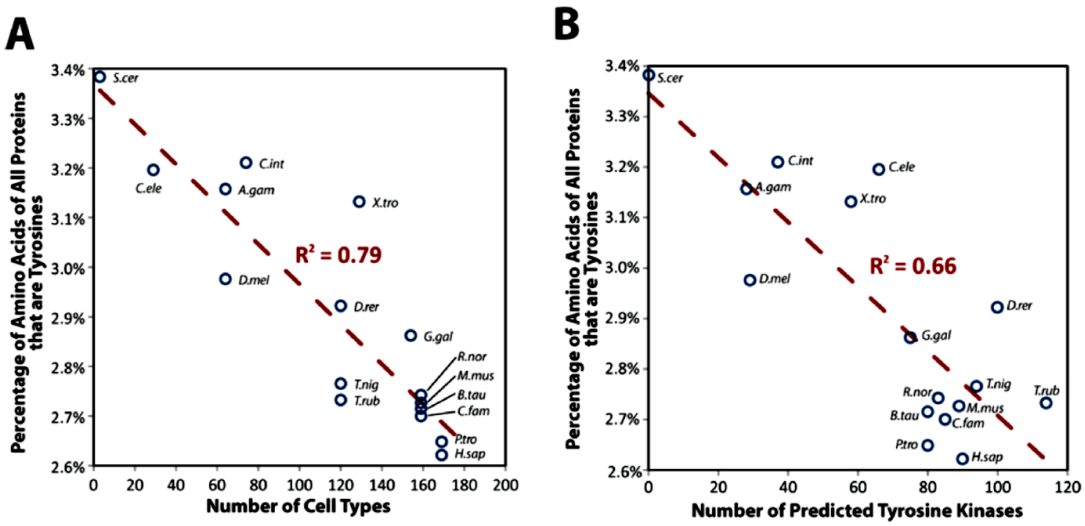


Figure 11: (A) The correlation between the evolution of multicellularity (emergence of different cell types) and the percentage of tyrosine residues within proteins. As the number of cell types increases, the percentage representation of tyrosine residues within proteins decreases. (B) The correlation between the number of predicted tyrosine kinases and the percentage of tyrosine residues within proteins. As the number of predicted tyrosine kinases increases, the percentage representation of tyrosine residues within proteins decreases. Distinct species were used (yeast (*S. cerevisiae*), worm (*C. elegans*), sea squirt (*C. intestinalis*), fly (*D. melanogaster*), mosquito (*A. gambiae*), zebrafish (*D. rerio*), tetraodon pufferfish (*T. nigroviridis*), Japanese pufferfish (*T. rubripes*), frog (*X. tropicalis*), mouse (*M. musculus*), rat (*R. norvegicus*), chicken (*G. gallus*), dog (*C. familiaris*), cow (*B. taurus*), chimpanzee (*P. troglodytes*) and human (*H. sapiens*). Adopted from 'Positive Selection of Tyrosine Loss in Metazoan Evolution' (Tan et al., 2009).

A study dealing with the functional constraints on phosphoproteomes was conducted in 2009 by Landry et al. They analyzed 2 099 yeast and 2347 human groups of orthologous proteins and reconstructed the ancestral sequences for the yeast and vertebrate lineages. They found out that phosphosites without known functions were more rarely phosphorylated than phosphosites with a function. Besides, phosphosites with a specific function evolved on average significantly more slowly than those without a function and phosphosites in known motifs were more conserved in disordered regions. However, phosphosites within disordered regions evolved generally faster than those within ordered regions, because a large number of phosphosites within disordered regions was non-functional (Landry et al., 2009). This highlighted the need of more structural data in determination of phosphosites.

Chen et al. studied evolutionary rates of phosphosites in 3 526 human–mouse–dog–opossum orthologous phosphoprotein groups (Fig. 12). They confirmed that phosphosites occur more frequently in disordered regions and on the protein surface. Furthermore, they found that serine and threonine phosphosites evolve more slowly than non-phosphorylated sites within not only disordered, but also ordered protein regions. Tyrosine phosphosites had similar evolution rate as non-phosphorylated tyrosine residues (Chen et al., 2010).

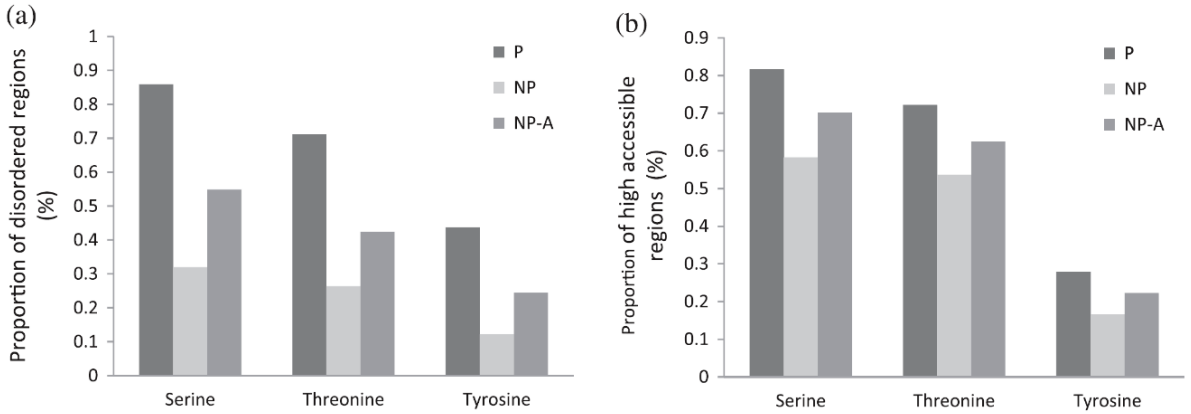


Figure 12: Comparison between phosphosites (P) and non-phosphorylated sites (NP, NP-A). Negative datasets were created by prediction (NP) and by experiment (NP-A). The proportion of phosphosites and non-phosphorylated sites localized within (A) disordered regions, (B) high accessible regions (on the surface or in cavities). Phosphosites were located more often within disordered and accessible regions, pSer and pThr significantly. Adapted from 'Phosphorylated and Nonphosphorylated Serine and Threonine Residues Evolve at Different Rates in Mammals' (Chen et al., 2010)

Trost et al. showed in 2016 that non-phosphorylated residues (serine, threonine, or tyrosine) were in their study more conserved than phosphosites. Protein kinases were more conserved than general proteins. The study was conducted on experimentally determined phosphosites from PhosphoSitePlus (Hornbeck et al., 2015) and Phospho.ELM databases (Dinkel et al., 2011) including phosphosites of mammals (*Homo sapiens*, *Pan troglodytes*, *Canis lupus familiaris*, *Mus musculus*, and *Rattus norvegicus*), insects (*Drosophila melanogaster*, *Apis mellifera*, and *Anopheles gambiae*), fish (*Danio rerio*, *Tetraodon nigroviridis*), plants (*Arabidopsis thaliana*, *Oryza sativa*, and *Ricinus communis*), birds (*Gallus gallus*), arachnids (*Ixodes scapularis*), nematodes (*Caenorhabditis elegans*), and others (*Plasmodium falciparum*, *Saccharomyces cerevisiae*, *Chlamydomonas reinhardtii*, and *Trypanosoma vivax*). The analysis of phosphosites conservation profiles further revealed that the degree of phosphosite conservation differed depending on the type of residues – serine residues were less conserved than threonine, or tyrosine and threonine residues were more conserved than serine, or tyrosine. These findings suggested that a variation in phosphosites can contribute to the variety of organisms phenotype (Trost et al., 2016).

In 2018, Miao et al. analyzed the conservation of phosphosites using 115 780 human phosphosites and 42 244 mouse phosphosites. They first reconstructed the ancestral sequences of the phosphorylation

proteins across 8 vertebrates and studied the appearance time of the phosphor accepting amino acids in the vertebrate species trees. The phosphosites were then divided into three groups based on the time of appearance: old, median and young. The old ones were the phosphosites that emerged earlier than 435 million years ago, the young ones those that have emerged since 96 million years ago. A median phosphosite was defined as one emerged between 96 and 435 million years ago. Miao et al. suggested that disordered regions had a high turnover rate (similarly to the hypothesis of (Landry et al., 2009)) and, therefore, many new phosphosites emerged there. However, most of them were non-functional and have then been eliminated by selection. So, the remaining phosphosites were important and therefore contributing to phenotypic fitness or neutral. Disordered regions could have been essential in the evolution of the phenotype differences (Miao et al., 2018).

Recently (the following information is from a preprint), the analysis of 537 321 phosphorylation sites (phosphosites) in 40 eukaryotic species, including 11 animals, 19 fungi, 7 plants and 3 apicomplexa species revealed that phosphosites were enriched in regulatory regions, at protein interfaces and near catalytic residues. Furthermore, 241 phosphosites were identified across a diverse set of 344 domains, which suggested a widespread ancient role of phosphorylation in regulation (Strumillo et al., 2018).

2.4 Protein phosphorylation sites prediction approaches

Protein phosphosites prediction tools are trained with experimentally annotated and known phosphosites to predict potential target sequences of protein kinases and thus reduce the number of sequences needed to be verified by e.g. mass spectrometry (Jensen, 2004).

Prediction tools can be based on sequence information such as the representation and distribution (order) of amino acids sequentially neighboring phosphosites, or they can focus on the properties of the three-dimensional (3D) environment of phosphosites, where the contribution of sequentially distant, but spatially close amino acids is relevant.

The protein phosphosites prediction tools can be protein kinase-specific or general. The kinase-specific approach requires as input both protein sequences and protein kinases and as output they provide the likelihood that each serine, threonine, and tyrosine in the sequence is phosphorylated by the chosen kinase. The general approach requires only protein sequences as input and reports the likelihood that each Ser, Thr and Tyr is phosphorylated by any possible kinase. The knowledge of kinases preferences can be used also for the training of general protein phosphosites prediction tools which leads to more precise prediction of phosphosites of known protein kinases. However, phosphosites with unknown protein kinases would be underestimated (Trost and Kusalik, 2011). In addition, the source of phosphorylated proteins can be taken into consideration in the prediction of phosphosites and these so called organism-specific approaches may achieve better performance accuracy (Trost and Kusalik, 2013).

2.4.1 Protein phosphorylation sites prediction tools

Protein phosphosites prediction tools may use simple machine learning methods and techniques such as Position-Specific Scoring Matrices (PSSM), Conditional random fields (CRFs), Random Forest (RF), Bayesian decision theory (BDT), and Hidden Markov Model (HMM), or more advanced ones such as artificial neural networks (ANNs), or support vector machine (SVM).

Position-Specific Scoring Matrices (PSSM) in their simplest form are matrices containing the frequencies of amino acids at given positions. PSSM are easy to construct and interpret, but cannot detect patterns in which combinations of amino acids are important (Trost and Kusalik, 2011). Several popular methods using this approach are presented in the following table 1.

Table 1: A list of protein phosphosites prediction tools using position-specific scoring matrices method. The source of training data (column named 'database'), the number of phosphosites used for training (pSer, pThr, pTyr, and pHis), reference, and website address are presented.

Prediction tools	Database	Phosphosites				References	Website
		pSer	pThr	pTyr	pHis		
PHOSITE	PhosphoBase	-	-	-	-	(Koenig and Grabe, 2004)	-
SMALI	SwissProt	-	-	-	-	(Li et al., 2008)	-
PAAS	Phospho.ELM	-	-	-	-	(Sobolev et al., 2010)	-
PostMod	Phospho.ELM	-	-	-	-	(Jung et al., 2010)	pbil.kaist.ac.kr/PostMod
-	NCBI RefSeq, P^3DB	2 444	635	268	-	(Huang et al., 2015)	-
PPRED	Phospho.ELM	13 320	2 766	2 166	-	(Biswas et al., 2010)	http://biomecis.uta.edu/~ashis/res/ppred/index.php
Quokka	Phospho.Elm UniProt	-	-	-	-	(Li et al., 2018)	http://quokka.erc.monash.edu/

- PhosphoBase (Blom et al., 1998), SwissProt (Bairoch and Apweiler, 2000), Phospho.ELM (Dinkel et al., 2011), NCBI RefSeq (Brister et al., 2015), P^3DB (Yao et al., 2014), UniProt (Bateman et al., 2017)

Conditional random fields (CRFs) is a method specifying the probabilities of possible label sequences given the observation sequence that can be past, present, or future. (Lafferty et al., 2001). Conditional random field was used in CRPhos. Dang et al. trained the prediction tools on 1 538 phosphosites across pSer, pThr, pTyr, and phosphorylated histidine obtained from Phospho.ELM (Diella et al., 2008) (Dang et al., 2008).

Random Forest (or Random Decision Forest) constructs a multitude of decision trees, each of which is built using a number of randomly selected features. The output is the class that is the mode (the most frequent value) of the classes (classification) or mean prediction (regression) of the individual trees. The more trees predict a given site as a phosphorylation site, the more likely this site is a positive one (Trost and Kusalik, 2013). Several popular methods using this approach are presented in the following table 2.

Table 2: A list of protein phosphosites prediction tools using Random Forest method. The source of training data (in column named 'database'), the number of phosphosites used for training (pSer, pThr, pTyr, and pHis), reference, and website address are presented.

Prediction tools	Database	Phosphosites				References	Website
		pSer	pThr	pTyr	pHis		
PHOSFER	Phospho.ELM PhosphoSite-Plus P^{3}DB PhosphoGRID	41 840	10 701	10 044	-	(Trost and Kusalik, 2013)	http://saphire.usask.ca/saphire/phosfer/index.html
Phospho-Predict	Phospho.ELM	1 853				(Akutsu et al., 2017)	http://phosphopredict.erc.monash.edu/

- Phospho.ELM (Dinkel et al., 2011), PhosphoSitePlus (Hornbeck et al., 2012), P3DB (Yao et al., 2014), PhosphoGRID (Stark et al., 2010)

Bayesian decision theory is a statistical system that tries to quantify the tradeoff between various decisions, making use of probabilities and costs. The probability distribution of each phosphosite is estimated and the Bayes risk for either potential solution (true or false phosphosite) is calculated, respectively. When protein kinase specific approach is employed, a difference profile of Bayesian decision risk is built for each protein kinase group in prediction. The decision whether a phosphosite is positive or negative depends on the level of difference between the prior and posterior probability of phosphorylation (Xue et al., 2006). Several popular methods using this approach are presented in the following table 3.

Table 3: A list of protein phosphosites prediction tools using Bayesian decision theory. The source of training data (column named 'database'), the number of phosphosites used for training (pSer, pThr, pTyr, and pHis), reference, and website address are presented.

Prediction tools	Database	Phosphosites				References	Website
		pSer	pThr	pTyr	pHis		
PPSP	Phospho.ELM	-	-	-	-	(Xue et al., 2006)	http://ppsp.biocuckoo.org/
IEPP	Phospho.ELM	396		39	-	(Wang et al., 2008)	-
Phospho-Pick	Phospho.ELM HPRD	2 964				(Kobe et al., 2015)	http://bioinf.scmb.uq.edu.au/phosphopick/phosphopick

- Phospho.ELM (Diella et al., 2004), HPRD (Mishra et al., 2006)

Hidden Markov Model (HMM) describes a probability distribution over a potentially infinite number of states. The state of a sequence through which the model is passing is hidden,wheras the output dependent on the state is visible. Therefore, a sequence of outputs generated by an HMM gives some information about the sequence of states (Wang et al., 2008). HMM was used for example in PREDIKIN (Brinkworth et al., 2003) and PREDIKIN2.0 (Ellis and Kobe, 2011). PREDIKIN was trained on data from PhosphoBase (Blom et al., 1998), PREDIKIN 2.0 on data from Phospho.ELM (Diella et al., 2004) and SwissProt (Bairoch and Apweiler, 2000).

Artificial neural networks are computing systems vaguely inspired by (the) biological neural networks that constitute animal brains. ANNs recognize the patterns seen during training and are then able to recognize similar, non-identical patterns. ANNs use a network of neurons, where each neuron has multiple inputs and a single output based on the weights associated with the various inputs. The output of each artificial neuron is sigmoidal in character, resulting in values from 0 to 1. The outputs are probabilities of the input data belonging to a certain category, so it can be used for classification. The neurons are organized in layers, each neuron is connected to every neuron in the next layer. The connections are weighed and initialized weights are gradually adjusted during the learning process (Berry et al., 2004). This increases output accuracy in each step and also well suppresses 'outliers' in the input data, highly atypical examples which may lead to false positives (Blom et al., 2004). However, ANNs together with SVM are difficult to interpret and suffer from enormous complexity making them CPU intensive (Trost and Kusalik, 2011). Several popular methods using this approach are presented in the following table 4.

Table 4: A list of protein phosphosites prediction tools using artificial neural networks. The source of training data (in column named 'database'), the number of phosphosites used for training (pSer, pThr, pTyr, and pHis), reference, and website address are presented.

Prediction tools	Database	Phosphosites				References	Website
		pSer	pThr	pTyr	pHis		
NetPhos	PhosphoBase	584	108	210	-	(Blom et al., 1999)	http://www.cbs.dtu.dk/services/NetPhos/
NetPhosK	PhosphoBase SwissProt	-	-	-	-	(Blom et al., 2004)	http://www.cbs.dtu.dk/services/NetPhosK/
NetPhos-Yeast	SwissProt	953	192	-	-	(Ingrell et al., 2007)	cbs.dtu.dk/services/NetPhosYeast
GANNPhos	Phospho.ELM	2 546	643	944	-	(Tang et al., 2007)	-
GPS 2.0	Phospho.ELM	9 717	1 818	1 719	-	(Xue et al., 2008)	http://www.gps.biocuckoo.org/
NetPhosBac	PHOSIDA	187			-	(Miller et al., 2009)	http://www.cbs.dtu.dk/services/
MusiteDeep	SwissProt UniProt RegPhos	34 401[a]		1 883	-	(Wang et al., 2017)	-
		1 827[b]		-	-		

(a) Training data set for general part of MusiteDeep
(b) Training data set for protein kinase-specific part of MusiteDeep

- PhosphoBase (Blom et al., 1998), SwissProt (Bairoch and Apweiler, 2000), Phospho.ELM (Dinkel et al., 2011), PHOSIDA (Gnad et al., 2007), UniProt (Bateman et al., 2017), RegPhos (Huang et al., 2014)

Support Vector Machine approach searches and studies general types of relations (e.g. clusters, correlations, classifications) in datasets. Heterogeneous biological data are represented by feature vectors in space. SVM tries to separate a given set of feature vectors with an optimal hyperplane. For this purpose, SVM uses a transformation into a dimension that has a clear dividing margin between categories. The optimum is reached when feature vectors are divided by a clear gap that is as wide as possible using a small number of support vectors. SVM has many advantages: it is easier to implement than neural networks, it can work successfully with a low number of observations when the feature

vectors are sparse, and avoids over-fitting by seeking a globally optimized solution, so a large number of features is permitted (Plewczyński et al., 2005). Several popular methods using this approach are presented in table 5.

Table 5: A list of protein phosphosites prediction tools using Support Vector Machine. The source of training data (in column named 'database'), the number of phosphosites used for training (pSer, pThr, pTyr, and pHis), reference, and website address are presented.

Prediction tools	Database	Phosphosites				References	Website
		pSer	pThr	pTyr	pHis		
Predphospho	PhosphoBase SwissProt	667-855	163-216			(Kim et al., 2004)	-
KinasePhos 2.0	Phospho.ELM SwissProt	11 888	2 433	2 179	43	(Wong et al., 2007)	http://kinasephos2.mbc.nctu.edu.tw/
PHOSIDA	PHOSIDA	4 731	664	107	-	(Gnad et al., 2007)	http://141.61.102.18/phosida/index.aspx
AutoMotif Server	SwissProt	12 103				(Plewczynski et al., 2008)	-
SiteSeek	PDB SwissProt Phospho3D Phospho.ELM	-				(Yoo et al., 2008)	-
PhosPhAt	PhosPhAt	802	-	-	-	(Heazlewood et al., 2008)	http://phosphat.uni-hohenheim.de/
Musite 1.0	Phospho.ELM SwissProt PhosphoPep PhosPhAt	61 448	14 478	5 727	-	(Gao et al., 2010)	http://musite.sourceforge.net/
PhosTryp	-	966	210	-	-	(Zilberstein et al., 2011)	http://phostryp.bio.uniroma2.it/
CKSAPP_PhSite	Phospho.ELM	12 373	2 525	1 826	-	(Zhao et al., 2012)	-
PTMpred	Phospho.ELM	1 694				(Xu et al., 2014)	http://doc.aporc.org/wiki/PTMPred
Phospho-SVM	Phospho.ELM	26 401	7 371	2 839	-	(Dou et al., 2014)	http://sysbio.unl.edu/PhosphoSVM/prediction.php
-	dbPTM Phospho.ELM PhosphoSite-Plus dbOGAP36 SysPTM	2 990	1 961	1 791	-	(Wang et al., 2015)	-
PredPhos	Phospho.ELM PhosphoPOINT PhosphoSite-Plus	2 404				(Gao et al., 2016)	-
KSRPred	Phospho.ELM PhosphoSite-Plus	6 839				(Wang et al., 2017b)	-
PhosContext 2vec	Phospho.Elm UniProt	7 021	2 515	2 066	-	(Xu et al., 2018)	http://phoscontext2vec.erc.monash.edu/

- PhosphoBase (Blom et al., 1998), SwissProt (Bairoch and Apweiler, 2000), Phospho.ELM (Dinkel et al., 2011), PHOSIDA (Gnad et al., 2007), PDB (Berman et al., 2000), Phospho3D (Zanzoni et al., 2011), PhosphoPep (Zanzoni et al., 2011), PhosPhAt (Heazlewood et al., 2008), dbPTM (Lee et al., 2006), dbOGAP36 (Wang et al., 2011), SysPTM (Li et al., 2009), PhosphoPOINT (Yang et al., 2008), UniProt (Bateman et al., 2017)

A performance comparison between protein phosphorylation prediction tools is difficult to create while protein phosphorylation prediction tools cannot be simply compared. The reasons are as follows: different versions of databases used in literature generate more or less bias during comparison, cross-validations for the same dataset are difficult because of the unavailability of trainable versions for most models and using the same independent dataset for testing all methods can be biased if some testing data have already been used as training data. The identification of a suitable testing dataset is therefore very difficult (Dang et al., 2008). However, a phosphorylation prediction tools performance comparison based on an independent dataset was done for kinase-specific protein phosphosites prediction tools. This is presented in the following table 6 from (Gao et al., 2016). The authors highlighted the accuracy and sensitivity of PredPhos developed by them. On the other hand, this performance comparison showed that common kinase-specific protein phosphosites prediction tools have achieved poor accuracy and specificity and so the adding of structural information could be beneficial.

Table 6: Performance comparison of PPSP (Xue et al., 2006), Kinasephos (Wong et al., 2007), NetPhosK (Blom et al., 2004), GPS (Xue et al., 2008), and PredPhos (Gao et al., 2016) based on an independent test dataset. Performance comparison table is from (Gao et al., 2016). The performance of each model was measured by six metrics: sensitivity ('Sn'), specificity ('Sp'), precision ('Pre') correlation coefficient ('CC'), and F1-score ('F1').

Tools	Kinase family	Sn	Sp	Pre	CC	F1
PPSP	PKA	1.000	0.540	0.096	0.228	0.176
	PKC	0.400	0.527	0.031	−0.028	0.058
	CK2	0.500	0.390	0.038	−0.047	0.071
	SRC	0.538	0.859	0.286	0.304	0.373
	MAPK	0.571	0.380	0.043	−0.021	0.081
Kinasephos	PKA	0.125	0.877	0.048	0.001	0.069
	PKC	0.200	0.863	0.053	0.034	0.083
	CK2	0.500	0.976	0.500	0.476	0.500
	SRC	0.115	0.960	0.231	0.103	0.154
	MAPK	0.571	0.937	0.308	0.381	0.400
NetphosK	PKA	0.375	0.914	0.176	0.204	0.240
	PKC	0.200	0.802	0.037	0.001	0.063
	CK2	0.500	0.805	0.111	0.158	0.182
	SRC	0.038	1.000	1.000	0.187	0.074
	MAPK	0.286	0.979	0.400	0.311	0.333
GPS	PKA	0.500	0.871	0.160	0.222	0.242
	PKC	0.600	0.695	0.070	0.119	0.125
	CK2	0.500	0.854	0.143	0.202	0.222
	SRC	0.462	0.871	0.273	0.265	0.343
	MAPK	0.571	0.789	0.118	0.182	0.195
PredPhos	PKA	0.571	0.779	0.100	0.164	0.170
	PKC	0.824	0.870	0.452	0.544	0.583
	CK2	1.000	0.659	0.176	0.341	0.300
	SRC	0.789	0.802	0.234	0.356	0.361
	MAPK	0.375	0.986	0.600	0.452	0.462

Protein phosphosites prediction tools including structural information of phosphosites are still few. For example, of the presented tools, DISPHOS (Iakoucheva et al., 2004), PHOSIDA (Gnad et al., 2007), Musite (Gao et al., 2010), or PhosTryp (Zilberstein et al., 2011) employ the knowledge of 3D phosphosites properties. DISPHOS and Musite use only disorder information (Iakoucheva et al., 2004)(Gao et al., 2010). PhosTryp added to disorder information also the protein secondary structure

(Zilberstein et al., 2011). PHOSIDA studied not only accessibility and structural flexibility, but also conservation of phosphosites (Gnad et al., 2007).

To sum up, recently published papers show that phosphosites exhibit likelihood to be located within disordered regions of proteins, in hinges, and loops. However, a generally high number of phosphosites indicates that the number of phosphosites within structured regions may not be negligible. On condition that only 10 % of all phosphosites reside within rigid secondary structure elements as proposed by (Tyanova et al., 2013) and (Jiménez et al., 2007), there still would be about 23 000 phosphosites in the human proteome suitable for structural analysis (Vlastaridis et al., 2017).

Phosphosites are predominantly present on the protein surface, exposed to the solvent. They are sequentially as well as spatially surrounded mostly with positively or negatively charged and hydrophilic amino acid residues (phosphosites are depleted with neutrally charged and hydrophobic residues). Recently conducted analyses of phosphosites as well as developed protein phosphosites predictions focused mainly on the recognition of sequentially dependent features of phosphosites.

3. Goals

Phosphorylation is the most frequent post-translational protein modification. There are estimated about 13 000 phosphoproteins and 230 000 phosphosites in the human proteome (Vlastaridis et al., 2017). Even though previous studies suggest that the majority of phosphosites are located within disordered regions of proteins, the number of phosphosites in well-structured regions of proteins may not be negligible. It is therefore justified to study phosphorylation sites in ordered domains. Furthermore, there is an increasing amount of 3D experimental data of phosphorylated phosphorylation sites in the structural databases.

The first aim of this thesis was to collect available structural data of phosphorylated phosphorylation sites and prepare a high-quality non-redundant dataset of phosphorylated phosphorylation sites.

The main aim then was to establish a structural characterization of phosphorylated phosphosites from the point of view of protein secondary structure, compactness, solvent accessibility, hydrophobicity, charge, and evolutionary conservation.

The next aim of this thesis was to find out to what extent these characteristics are similar to the results obtained from non-phosphorylated phosphosites and to determine whether phosphorylated phosphosites may be useful in prospective protein phosphosites predictions.

The last aim of this thesis was to find how often phosphorylation leads to big conformational changes in protein structures.

4. Methods

4.1 Datasets

To study the properties of phosphosites three datasets were created: First dataset (pP-set) contains phosphorylated phosphosites, second dataset (wP-set) contains non-phosphorylated phosphosites, and third dataset (NonP-sets) contains non-phosphorylated residues of the same type as the phosphorylated residue (e.g. serine – Ser in case of pSer) within the protein structure containing phosphorylated phosphosites (experienced same phosphorylation conditions).

To get the first dataset (pP-set), protein structures containing phosphorylated phosphosites were obtained from the PDBeChem database (Dimitropoulos et al., 2006) for each amino acid (Ser, Thr and Tyr). There were 1031, 869, and 566 protein structures for pSer, pThr and pTyr by 9. 7. 2018.

To get the second dataset (wP-set), the data from the first dataset (pP-set) were used. For each protein from the pP-set, the list of all other protein structures of this protein (wP-set) was obtained using Structure Integration with Function, Taxonomy and Sequence (SIFTS) files (Bateman et al., 2017)(Velankar et al., 2013). SIFTS facilitates mapping between UniProt and PDB. A script was written to download SIFTS and extract information using modified Python package 'BeautifulSoup'. When more than one structure was found for each phosphorylated phosphosites, the one with highest sequence identity was selected.

To get the third dataset (nonP-set), non-phosphorylated residues of the same type as phosphosites (e.g. serine in case of pSer) were selected from the protein structures within pP-set (containing phosphosite(s)).

Protein structure data were then downloaded from Protein Data Base (Berman et al., 2000) and processed by own script: structures were filtered to reduce a redundancy, to avoid peptides and to select high quality structures as described below.

4.1.1 Elimination of peptides

Protein structures with a chain length less than 50 amino acids (aa's) typically represent peptides or not fully folded domain. Therefore, structures with the chain length smaller than 50 aa's were studied manually (Supplementary material, Table 1). Protein structures with the chain length less than 30 aa's were all not folded properly or represented peptides. Protein structures with the chain length between 30 and 50 aa's had proper-folded domain in 5 cases (2ljd, 2lje, 2ljf, 2joc and 2mx4). Protein structures without proper-folded domain or represented peptides (641) were excluded from analyses.

4.1.2 R-value, R-free and resolution

To include only high-quality protein structures, protein structures were filtered based on R-value and resolution. R-value (or R-work, R-factor) is a measure of the agreement between the crystallographic model and the experimental X-ray diffraction data. R-value is defined as:

$$R = \frac{\sum\left||F_{obs}| - |F_{calc}|\right|}{|F_{obs}|}$$

where F_{obs} is structure factor observed by experiment and F_{calc} is structure factor calculated within the crystallographic model.

Therefore, low R-value refers to highly matched diffraction pattern to the experimentally-observed diffraction pattern. A theoretical perfect fit would have a value of 0. R-free value is calculated by seeing how well the model predicts the 10% that were not used previously to determine R-value. R-free is typically a little bit higher than R-value (J.Kleywegt and Jones, 1997).

Resolution is a measure of the distance corresponding to the smallest observable detail present in the diffraction pattern and the detail that will be seen when the electron density map is calculated. Resolution value presented in PDB file is a median value for the whole protein structure (Wlodawer et al., 2008).

R-values and resolution values were extracted from PDB files and protein structures with R-values more than 0.3 and with resolution higher than 3.5 were filtered out from datasets using own script.

4.1.3 Redundancy

Redundancy of datasets was eliminated by CD-HIT (Huang et al., 2010). For each dataset, two subsets (30 and 90) were made, where the number corresponds to the maximum sequence identity shared by the protein chains in the redundancy set. These subsets were compared with subsets obtained from PISCES tool (Wang and Dunbrack, 2005) for the same parameters (resolution ≤ 3.5, R-value ≤ 0.3 and sequence identity 30/90). The number of phosphosites was equal in both.

To sum up, three nonredundant subsets were made for each amino acid (Ser, Thr and Tyr): first one contains phosphorylated phosphosites (pP-sets), second one contains phosphosites that were identified as phosphorylated in homologous protein structure(s) but within chosen structure they were non-phosphorylated (wP-sets), and third one contains non-phosphorylated residues within the protein structure including phosphorylated residue(s) (NonP-sets). Because for each amino acid the maximum sequence identity shared by the protein chains in the redundancy set was defined as 30 % and 90 %, 18 subsets were made in total (Fig. 13).

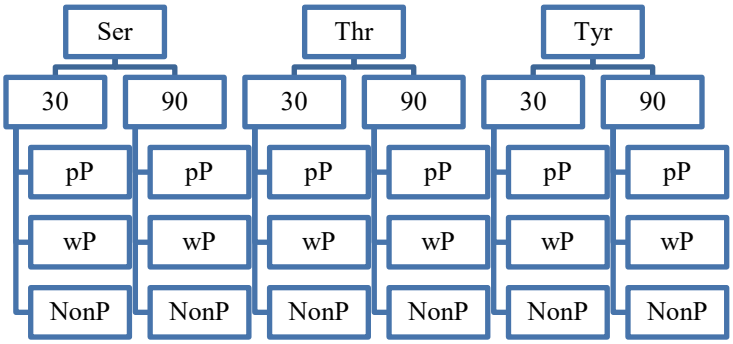


Figure 13: Diagram of datasets used in analyses. For each amino acid (serine - Ser, threonine - Thr, and tyrosine –Tyr) two subsets (30 and 90) were made, where the number corresponds to the maximum sequence identity shared by the protein chains in the redundancy set. For both 30 and 90 subsets three subsets were made based on the phosphorylation state of the protein site: 'pP-sets' contains phosphorylated phosphosites, 'wP-sets' contains non-phosphorylated phosphosites, and 'NonP-sets' contains non-phosphorylated residues of the same type as in pP-sets (serine in the case of pSer, threonine in pThr, and tyrosine in pTyr) within the protein structure containing phosphorylated residues.

4.2 Features extraction

4.2.1 Definition of phosphorylation sites surroundings

For the analysis of protein phosphosites 3D environmental properties is necessary to define which amino acids are in spatial neighborhood of the phosphosite. Two different methods were used, Euclidean distance and Voronoi diagrams (Edelsbrunner and Seidel, 1986), in order to describe the neighborhood complexly.

4.2.1.1 Euclidean distance

The Carbon serving as an acceptor of the phosphate was used as a center of a sphere. The sphere further defines the atoms located in spatial neighborhood of the phosphosite. Amino acid was selected as spatially neighboring the phosphosite if the distance between the Carbon of phosphosite and at least one atom of that amino acid was equal or smaller than the distance defined by Euclidean distance. Euclidean distance is defined as:

$$d_{(p,q)} = d_{(q,p)} = \sqrt{(q_x - p_x)^2 + (q_y - p_y)^2 + (q_z - p_z)^2}$$

where q is a Carbon and p is any atom in the surroundings. Each atom in crystal protein structure is descripted by coordinates x, y, and z. Carbons served as a center of the sphere are annotated in PDB files as OG1, OG and OH for pThr, pSer and pTyr, respectively.

The maximum radius of the sphere was set to 0, 65 nm (6.5 Å). Within this range there could be interaction among atoms (Hou et al., 2018).

4.2.1.2 Voronoi diagrams

Voronoi diagram approach partitions a space into the regions based on distance to points in a specific subset of space. That set of points is specified beforehand (atoms of amino acids), and for each point there is a corresponding region consisting of all points closer to that point than to any other. Connection between every two points is made and a ridge between them is defined as the midpoint of this connection (Aurenhammer, 1991). Two residues are said to be neighbors when at least one pair of atoms of each residue have a common ridge. Voronoi diagrams were calculated using the Qhull python library called 'scipy.spatial.Voronoi' (Barber and Dobkin, 1996).

Both approaches for the definition and selection of aa's spatially neighboring the phosphosites are illustrated on 2D Voronoi diagrams in Fig. 14.

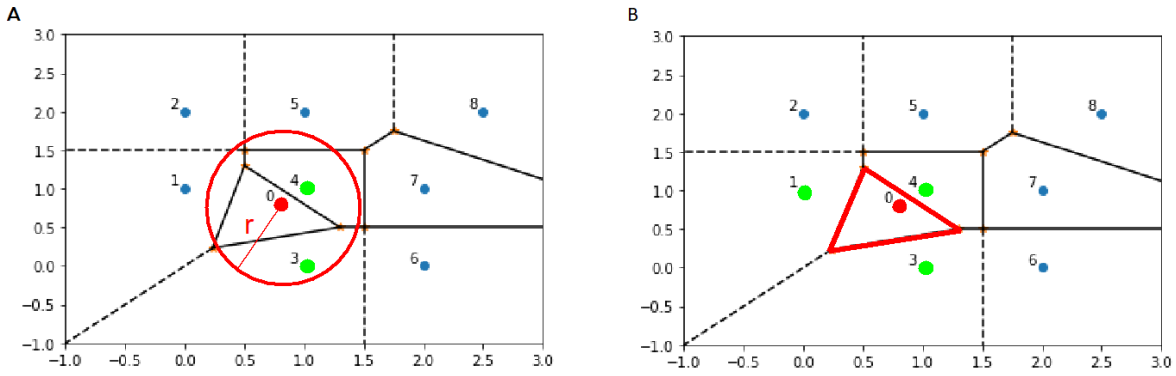


Figure 14: 2D representation of the definition of the amino acids spatially neighboring the phosphorylation site by two approaches: (A) Euclidean distance, and (B) Voronoi diagrams. Points represents amino acids. A red point is the phosphosite, in the Euclidean distance approach a center of a sphere too. Green points represents the amino acids selected by particular approach as neighboring the phosphosite, not selected amino acids are represented by blue points. Graph illustrates differences between (A) Euclidean distance and (B) Voronoi diagrams amino acids selection. Whereas (A) Euclidiean distance determined amino acids of the numbers 3 and 4 as the phosphosite neighbors, (B) Voronoi diagram approach determined amino acid of the number 1 as the phosphosite neighbor too. Graph was made using Python 'matplotlib.pyplot" package.

4.2.2 Features

4.2.2.1 Protein secondary structure

Previously published papers (e.g. (Tyanova et al., 2013) or (Karabulut and Frishman, 2016)) showed less frequent localization of phosphosites within structured elements, analysis of secondary structure states of phosphosites was therefore conducted. Protein secondary structure annotations were extracted from the EBI database Structure Integration with Function, Taxonomy and Sequence (SIFTS) files (Velankar et al., 2013). SIFTS protein secondary structure annotations of each amino acids within the protein structure was done by Define Secondary Structure of Proteins (DSSP) algorithm (Kabsch and Sander, 1983). Amino acids were annotated: 'helix', 'strand', 'loop'. Whether the phosphosite was annotated as acquiring two possibilities, it was defined as 'microheterogeneity'. Hetero atoms that have no protein secondary structure were annotated: 'HETATM'.

4.2.2.2 Compactness

Compactness of the surrounding of phosphosites suggests how much the amino acids interact with each other and tend to be spatially located close and fulfil the space. Compactness of the protein sites is defined through Euclidean Cα-Cα distances of spatially neighboring amino. Less compact sites would have lower average Cα-Cα distance (Konrat, 2009). Euclidean distance is defined as:

$$d_{(p,q)} = d_{(q,p)} = \sqrt{(q_x - p_x)^2 + (q_y - p_y)^2 + (q_z - p_z)^2}$$

where q is a Cα of a central amino acid (Ser, Thr, Tyr, pSer, pThr, and pTyr) and p is a Cα of the surrounding residue. Each atom in crystal protein structure is descripted by coordinates x, y, and z.

4.2.2.3 Surface accessibility

Previously published papers (e.g. (Durek et al., 2009) or (Frades et al., 2015)) showed that surface accessibility may be one of the properties that can be used to distinguish between phosphosites and non-phosphorylated residues. Phosphosites could be preferentially located on the protein surface to be accessible for protein kinases. Surface accessibility of certain amino acid can be predicted using a machine learning approach trained on the known values of surface accessibilities within solved protein structures. The value of an absolute surface area (ASA) for experimentally solved structures is given in $Å^2$ and 'the area is calculated by rolling a sphere the size of a water molecule over the protein surface' (quoted from (Petersen et al., 2009)). Relative surface area (RSA) is defined as:

$$RSA = \frac{ASA}{ASA_{max}} * 100$$

where ASA is observed ASA for each particular amino acid and ASA_{max} is calculated for each amino acid as a maximum obtainable solvent exposed area in a tri-peptide flanked with glycine or alanine (Petersen et al., 2009).

Relative and absolute surface accessibility of aa's within protein structures were predicted using NetSurfP-2.0 (Klausen et al.). It was recently developed as the second version of NetSurfP-1.0 (Petersen et al., 2009), and now it is described in preprint (therefore citation does not contain a year). Residues were defined as exposed when RSA was higher or equal to 25 %. Residues with RSA lower than 25 % were defined as buried.

4.2.2.4 Hydrophobicity

Hydrophobicity of the phosphosite and its spatially neighboring aa's suggests the solvent exposure of aa's within protein structures and therefore can be used as an additional approach to the solvent accessibility prediction tools. Hydrophobicity of aa's sequentially neighboring the phosphosites was previously studied by for example (Frades et al., 2015), or (Huang et al., 2015). Therefore,

hydrophobicity of aa's spatially neighboring phosphosites is potential additional feature to distinguish between phosphosites and non-phosphorylated residues.

Different hydrophobicity scales are used for the determination of amino acid hydrophobicity – based on physicochemical properties or on a frequency of certain amino acid to be located at protein surfaces. Hydrophobicity scales based on physicochemical properties were chosen to avoid redundant information in the analysis (the previous approach for the determination of solvent accessibility was based on a machine learning approach).

Hydrophobicity scales from (Wolfenden et al., 1979) and (Radzicka and Wolfenden, 1988) were compared. They differed in a hydrophobicity of threonine, serine and tyrosine. While Wolfenden et al. suggested that these aa's are hydrophobic, Radzicka and Wolfenden grouped them to hydrophilic aa's (Wolfenden et al., 1979) (Radzicka and Wolfenden, 1988). Therefore, own hydrophobic scale was made, where threonine, tyrosine and serine were defined as polar and other aa's were grouped by hydrophobicity based on hydrophobicity scales from (Wolfenden et al., 1979) and (Radzicka and Wolfenden, 1988). Classification of aa's into three groups ('hydrophobic', 'hydrophilic', and 'polar') is shown in table 7. Glycine, leucine, isoleucine, valine, alanine, phenylalanine, cysteine, methionine, and tryptophan were defined as hydrophobic. Aspartate, lysine, glutamine, glutamate, histidine, asparagine, and arginine were defined as hydrophilic.

Table 7: Classification of amino acids into three groups ('hydrophobic', 'hydrophilic', and 'polar') by hydrophobicity. Hydrophobicity scales from (Wolfenden et al., 1979) and (Radzicka and Wolfenden, 1988) were used.

Hydrophobicity	Amino acids
hydrophobic	glycine, leucine, isoleucine, valine, alanine, phenylalanine, cysteine, methionine, tryptophan, proline
hydrophilic	aspartate, lysine, glutamine, glutamate, histidine, asparagine, and arginine
polar	threonine, tyrosine, serine

4.2.2.5 Charge

Charge of aa's spatially neighboring the phosphosite suggests whether phosphosite tends to be located within positively, negatively, or neutrally charged three-dimensional space.

Amino acids were grouped by charge using rules from the book 'Introduction to Proteins: Structure, Function, and Motion' (Kessel and Ben-Tal, 2018) (table 9). Lysine, arginine, and histidine were defined as positively charged and aspartate and glutamate as negatively charged. Alanine, cysteine, phenylalanine, glycine, isoleucine, leucine, methionine, asparagine, proline, glutamine, serine, threonine, valine, tryptophan, tyrosine were defined as neutral (not bearing a charge). Heteroatoms were treated as "HETATM" and excluded from the analysis.

Table 9: Amino acids grouped by charge using rules from the book 'Introduction to Proteins: Structure, Function, and Motion' (Kessel and Ben-Tal, 2018). Amino acids were defined as positively charged (= 'positive'), negatively charged (= 'negative') or without a charge (= 'neutral').

Charge	Amino acids
positive	lysine, arginine, histidine
negative	aspartate, glutamate
neutral	alanine, cysteine, phenylalanine, glycine, isoleucine, leucine, methionine, asparagine, proline, glutamine, serine, threonine, valine, tryptophan, tyrosine

4.2.2.6 Evolutionary conservation profiles

Evolutionary conservation profiles of phosphosites and aa's spatially neighboring them suggest the functional role of phosphosites. The rate of evolution of each phosphosite can indicate for example, whether it is a functionally important site, or a randomly newly-emerged phosphosite without the specific function (Landry et al., 2009).

Evolutionary conservation profiles for each amino acid within the protein structure were obtained from ConSurf-DB (Goldenberg et al., 2009). The rate of evolution, and thus the conservation, at each site was estimated by Rate4Site. Firstly, Rate4Site constructed a phylogenetic tree. The rates of evolution were assumed to follow a Gamma distribution that was used as the prior in a Bayesian interference scheme (Brinkworth et al., 2008). The resulting conservation scores were normalized – an average score over all amino acids was zero, and the standard deviation was one. Negative values of conservation scores indicate the conserved positions while positive scores indicate the variables one.

ConSurf-DB then formed evolutionary conservation profiles using these scores. Profiles contained a scale of evolutionary conservation in a range (1-9), where 1 is the most variable one, 9 is the most conserved one. Rate4Site required multiple sequence alignment of at least 50 unique homologous. Therefore, several evolutionary conservation profiles of protein structures could not be created.

4.2.2.7 Structure comparison

Structures of pairs of phosphosites (in phosphorylated and unphosphorylated state) were aligned to study conformational changes in protein structures caused by phosphorylation. Therefore, root-mean-square deviation of atomic positions (RMSD) was calculated for phosphosites before and after the phosphorylation. RMSD capture the difference between conformations of the phosphosite through the paired protein structures from pP- and wP-sets that were superposed by Dali server (Holm and Laakso, 2016).

RMSD is defined as:

$$RMSD = \sqrt{\frac{1}{N} \times \sum_{i=1}^{N} \left(\left(a_{i_x} - b_{i_x}\right)^2 + \left(a_{i_y} - b_{i_y}\right)^2 + \left(a_{i_z} - b_{i_z}\right)^2 \right)}$$

Where 'N' is a sum of all atoms shared by amino acid before and after the phosphorylation, 'a' is an atom of the phosphosite used as a reference and 'b' is an atom in a new protein conformation (Coutsias et al., 2004). Each atom in crystal protein structure is descripted by coordinates x, y, and z. All atoms of the amino acids shared before and after the phosphorylation were superimposed and final RMSD represents the average of all RMSD values for these atoms (Fig. 15).

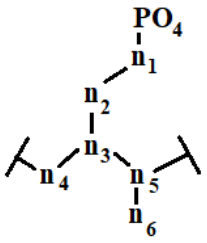


Figure 15: Scheme of serine residue from RMSD point of view. All atoms named 'n' were used for RMSD calculation. The average of distances between all atoms shared by amino acid before and after phosphorylation was measured.

4.3 Statistical methods

A statistical significance of results was measured by chi-squared test, analysis of variance, or single-sample t-test. These approaches are briefly presented below. Chi-squared test was used for the analyses of protein secondary structure states and solvent accessibility of phosphosites, as well as for the analyses of hydrophobicity and charge of aa's spatially neighboring the phosphosites. Analysis of variance (multiple) was used for the analyses of compactness and conservation of phosphosites and spatially neighboring aa's. Single-sample t-test was used for the analysis of how much the position of phosphosites side chains in space differs before and after phosphorylation.

4.3.1 Chi-squared test

Pearson's chi-squared test (χ^2) was used to test whether an observed frequency distribution fits the theoretical distribution (test of goodness of fit).χ^2 statistic was calculated using:

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - Np_i)^2}{Np_i}$$

Where N is a number of measurements that had k different outcomes. Frequency of outcome i was denoted by X_i. Zero hypothesis was defined as the probability of outcome i to be p_i. By comparing χ^2

with critical value from chi-squared distribution with df degrees of freedom and selected confidence level the null hypothesis was accepted or rejected. If χ^2 exceeded the critical value of chi-squared distribution, the null hypothesis was rejected with the selected level of confidence (Pearson, 1900).

Degrees of freedom (df) is number of categories (k) minus 1, reduced by the number of unknown parameters of the fitted distribution. The main bottleneck of this statistic test is a minimal frequency of outcomes. Although there is no strict mathematical assumption, the wide accepted rules say that no theoretical frequency (Np_i) can be zero, and at least 80 % of theoretical frequencies should be 5 or more (for contingency tables of size 2 × 2 even all the cells should have values of 5 or more). In this thesis, all cases obeyed the first rule and a majority obeyed the second rule. In several occasions the theoretical frequency was under limit value of 5. When the results are far enough from the critical value, limited accuracy is acceptable.

4.3.2 Analysis of variance

Analysis of variance (ANOVA) was used to determine differences in group means in a sample. The basic setting of ANOVA can only confirm or reject (with selected level of confidence) zero hypothesis saying, that all groups have the same means. Because we had more than two groups, more detailed analyses were necessary. It was not feasible to mutually compare pairs of groups, because of error accumulation. Multiple testing would result in a high probability of finding significant difference only by chance. The problem of multiple comparison was solved by means of Tukey's honestly significant difference (HSD) test (Tukey, 1949).

4.3.3 Student's t-test

Student's t-test (or simply t-test) is a statistical method for testing if the means on two sets of data differ significantly. Basic assumption states, that n observations are possessing a normal distribution and have an unknown variance σ^2. There are several variants of t-test, in this thesis one sampled paired t-test was used. Test null hypothesis stated that couples of values (y_i, z_i) had equal means (same as testing if mean of $x_i = y_i - z_i$ equals zero). Denoting average of x_i as $\overline{x}$ and variance of x_i as s^2, the t-test is defined as:

$$t = \frac{\overline{x}}{s}\sqrt{n}$$

't' was compared with critical value from Student's t-distribution with n-1 degrees of freedom ('df' - the number of parameters of the system that may vary independently) and selected confidence level. In addition to p-value, t-value was obtained which represents a size of the difference relative to the variation in sample data (Xu et al., 2017). If it exceeded the critical value of t-distribution, we rejected the null hypothesis with the selected level of confidence.

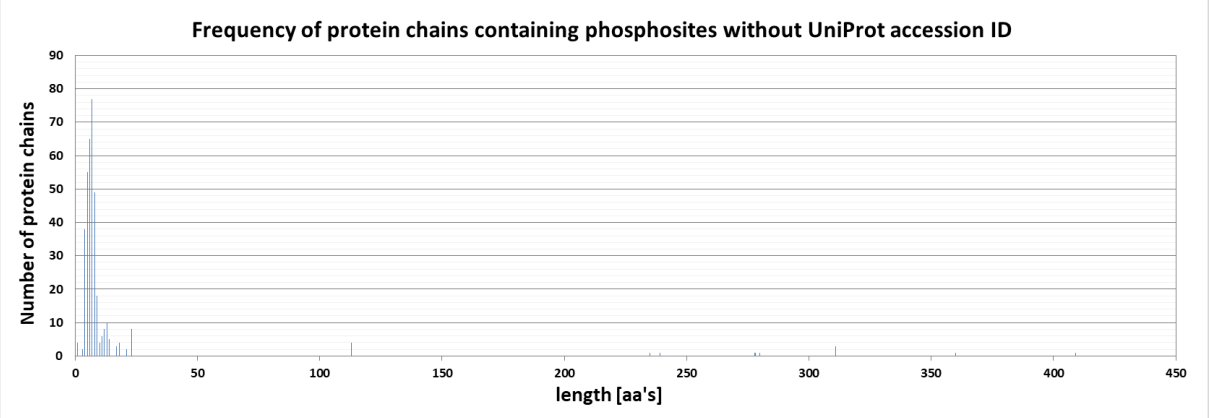
5. Results

5.1 Datasets characteristics

For each amino acid (Ser, Thr, Tyr) three subsets based on the characters of sites were made: ‘pP’ contained phosphorylated phosphosites, ‘wP’ contained non-phosphorylated phosphosites, and ‘NonP’ contained non-phosphorylated residues of the same type as phosphosite (serine in the case of pSer, threonine in pThr, and tyrosine in pTyr) (see 4.1). In the section below the term ‘sites’ would be used in the case of general description of all sites (pP-, wP-, as well as NonP-sites).

Phosphosites were filtered based on length and proper-folding of domain of protein chains, and resolution and R-values of protein structures. Protein chains with length less than 30 aa’s were excluded. This agrees with the UniProt politics not to collect peptides. UniProt is a database within each protein sequence has a unique accession ID (Bateman et al., 2017). Analysis of the presence of UniProt accession ID for the proteins in datasets showed the missing accession ID values mainly for protein chains less than 30 aa’s long (Fig. 16). All protein chains longer than 450 aa’s had an UniProt accession ID.

(A)



(B)

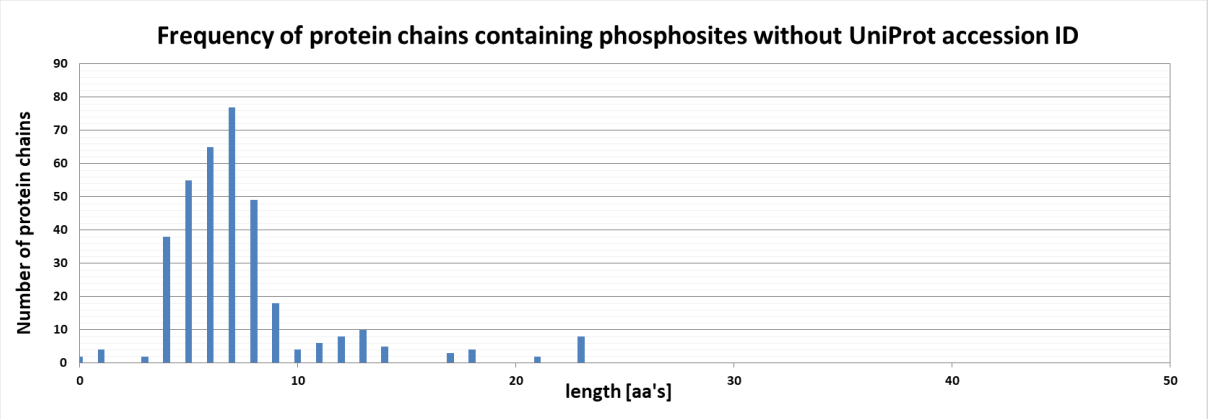


Figure 16: Frequency of protein chains containing phosphosites without UniProt accession ID. (A) Protein chains with length between 0 and 450 aa’s. (B) Protein chains with length between 0 and 50 aa’s.

The number of phosphosites and protein structures for each step is summed in table 10. At the beginning, pP-sets contained 1 263 pTyr, 2 097 pSer, and 1 544 pThr within 566, 1 031, and 869 protein structures, respectively. Folding of these protein structures containing pP-sites were studied and protein structures without proper-folded domain were excluded. After this step, 858 pTyr, 1 500 pSer, and 1 314 pThr remained within 350, 751, and 724 protein structures, respectively. Filter based on resolution followed and protein structures with resolution higher than 3.5 were excluded. After this step, 846 pTyr, 1 402 pSer, and 1 280 pThr remained within 347, 727, and 712 protein structures, respectively. Then, protein structures with R-value higher than 0.3 were excluded. After this step, 835 pTyr, 1 364 pSer, and 1 265 pThr remained within 337, 704, and 699 protein structures, respectively.

Table 10: The number of phosphosites and protein structures in pP-sets for each amino acid (Tyr, Ser, Thr). Without any filter ('raw'), after the elimination of peptides and protein structures without proper-folded domain, resolution and R-value filters.

	raw		proper-folded domain		resolution		R-value	
dataset	Protein structures	Sites	Protein structures	Sites	Protein structures	Sites	Protein structures	Sites
Tyr	566	1 263	350	858	347	846	337	835
Ser	1 031	2 097	751	1 500	727	1 402	704	1 364
Thr	869	1 544	724	1 314	712	1 280	699	1 265

Similar steps were followed also for wP-datasets with an extra step including the check whether on the position of phosphosites is the residue of same type as phosphosite in pP-set (e.g. serine in case of pSer). Several protein structures with identical UniProt accession ID contained mutated aa's in this position (for example aspartate and glutamate) as can be demonstrated by decreasing of number of wP-sites after this filtering step. The number of wP-sites and associated protein structures is presented in table 11. At the beginning, wP-sets contained 3 442 pTyr, 3 934 pSer, and 3 523 pThr within 1 989, 2 087, and 2 177 protein structures, respectively. After the check of amino acid type, 2 332 pTyr, 2 279 pSer, and 1 752 pThr remained within 1 498, 1 483, and 1 382 protein structures, respectively. After the filter based on protein folding, resolution, and R-value 1 766 pTyr, 1 433 pSer, and 975 pThr remained within 1 148, 942, and 753 protein structures, respectively.

Table 11: The number of phosphosites and protein structures in wP-sets for each amino acid (Tyr, Ser, Thr). Without any filter ('raw'), after the elimination of peptides and protein structures without proper-folded domain, resolution and R-value filters.

	raw		type		proper-folded domain, resolution, R-value	
dataset	Protein structures	Sites	Protein structures	Sites	Protein structures	Sites
Tyr	1 989	3 442	1 498	2 332	1 148	1 766
Ser	2 087	3 934	1 483	2 279	942	1 433
Thr	2 177	3 523	1 382	1 752	753	975

Finally, the redundancy was eliminated by CD-HIT (Huang et al., 2010). For each dataset two subsets (30 and 90) were created, where the number corresponds to the maximum sequence identity shared by the protein chains in the redundancy set. NonP-sites were then found within non-redundant protein structures containing pP-sites. The number of sites across Ser, Thr, and Tyr datasets is presented in table 12. Tyrosine pP-set contained 29 and 70 protein structures including 32 and 91 sites for Tyr30 and Tyr90, respectively. Serine pP-sets contained 94 and 134 protein structures including 117 and 166 sites for Ser30 and Ser90, respectively. Threonine pP-sets contained 56 and 105 protein structures including 64 and 119 sites for Thr30 and Thr90, respectively. Tyrosine wP-set contained 20 and 46 protein structures including 23 and 61 sites for Tyr30 and Tyr90, respectively. Serine wP-sets contained 52 and 77 protein structures including 63 and 92 sites for Ser30 and Ser90, respectively. Threonine wP-sets contained 31 and 48 protein structures including 40 and 57 sites for Thr30 and Thr90, respectively. Tyrosine NonP-set contained 29 and 70 protein structures including 423 and 880 sites for Tyr30 and Tyr90, respectively. Serine NonP-sets contained 94 and 134 protein structures including 1 819 and 2 485 sites for Ser30 and Ser90, respectively. Threonine NonP-sets contained 56 and 105 protein structures including 893 and 1 592 sites for Thr30 and Thr90, respectively.

Table 12: The number of protein structures and sites included in datasets for tyrosine, serine, and threonine. Three different types of sites were collected: phosphorylated phosphosites (pP), phosphosites that were not phosphorylated (wP), and non-phosphorylated sites (NonP). For each amino acid two sets (30 and 90) were made, where the number corresponds to the maximum sequence identity shared by the protein chains in the redundancy set.

	pP-sets		wP-sets		NonP-sets	
	Protein structures	sites	Protein structures	sites	Protein structures	sites
Tyr30	29	32	20	23	29	423
Tyr90	70	91	46	61	70	880
Ser30	94	117	52	63	94	1 819
Ser90	134	166	77	92	134	2 485
Thr30	56	64	31	40	56	893
Thr90	105	119	48	57	105	1 592

Further, pairs between pP- and wP-sets were made, where pairs have the same UniProt accession ID. To get a comprehensive list of wP-sites, pP-sites before the filtering steps were used as the template list of UniProt accession ID's. Several pP-sites that were used for the finding of wP-sites did not pass through the filter steps and the number of pairs are, therefore, lower than the number of wP-sites itself. Summary of the number of pairs is presented in table 13.

Table 13: The number of protein structures and sites included in paired datasets (pP-wP) for tyrosine, serine, and threonine. For each amino acid two sets (30 and 90) were made, where the number corresponds to the maximum sequence identity shared by the protein chains in the redundancy set.

	Protein structures	sites
Tyr30	18	18
Tyr90	44	55
Ser30	49	53
Ser90	66	73
Thr30	27	27
Thr90	40	41

Site can be described by its properties such as the localization on the protein surface/within the protein structure, within the protein secondary structure elements, evolutionary conservation level, conformational isomerism, or by the properties of the spatially surrounding aa's such as the hydrophobicity, charge or compactness.

5.2. Protein secondary structure

Protein secondary structure of site can partially response to the ordered/disordered state of the region containing this site. Preferences of pP-, wP-, and NonP-sites to reside in specific protein secondary structure element was studied.

Several sites were not annotated for unknown reason (one in Tyr30-pP, Tyr90-pP, Ser90-pP and Thr90-wP, and 41 protein sites in Tyr30-NonP, Tyr90-NonP, and 24 protein sites in Ser90-NonP). These not-annotated sites were together with the sites annotated as 'microheterogeneity' excluded from analysis. Results are shown in table 14 and the percentage distributions of pP-, wP- and NonP-sites in protein secondary structure elements are depicted in Fig. 17.

Chi-square test was used to evaluate a statistical significance of the differences between the distributions of pP- and wP-sites in protein secondary structure elements. Zero hypothesis ($H_{0)}$ were:

1. Distribution of pP- and wP-sites in protein secondary structure elements is random.
2. Distribution of pP- and NonP-sites in protein secondary structure elements is random.
3. Distribution of wP- and NonP-sites in protein secondary structure elements is random.

Results are presented in Supplementary material, table 2, and summarized in table 15.

Table 14: The distribution of pP-, wP- and NonP- sites in protein secondary structure elements. For serine = 'Ser', threonine = 'Thr', and tyrosine = 'Tyr' two sets (30 and 90) were made, where the number corresponds to the maximum sequence identity shared by the protein chains in the redundancy set. Three categories of protein secondary structures were used according to DSSP (Kabsch and Sander, 1983): 'helix', 'strand', and 'loop' Whether the site was annotated as acquiring two possibilities, it was defined as 'microheterogeneity' (= 'hetero'). Sites annotated as 'hetero' were not used in analysis (red colored). The distribution of sites in protein secondary structure is presented as the number ('count') and as a relative percentage within a dataset ('percent').

dataset	type	count				percent			
		helix	strand	loop	hetero	sum	helix	strand	loop
Tyr30	pP	7	4	20	1	31	22.580	12.903	64.516
	wP	10	2	11	0	23	43.478	8.695	47.826
	NonP	177	90	115	1	382	46.335	23.560	30.104
Tyr90	pP	10	26	54	1	90	11.111	28.888	60.000
	wP	14	11	36	0	61	22.950	18.032	59.016
	NonP	367	188	284	1	839	43.742	22.407	33.849
Ser30	pP	36	5	76	0	117	30.769	4.273	64.957
	wP	26	1	36	0	63	41.269	1.587	57.142
	NonP	856	242	721	0	1 819	47.058	13.304	39.637
Ser90	pP	51	6	108	1	165	30.909	3.636	65.454
	wP	35	1	56	0	92	38.043	1.086	60.869
	NonP	1 177	300	984	1	2 461	47.826	12.190	39.983
Thr30	pP	19	3	42	0	64	29.687	4.688	65.625
	wP	14	4	22	0	40	35.000	10.000	55.000
	NonP	349	165	379	0	893	39.081	18.477	42.441
Thr90	pP	30	4	85	0	119	25.210	3.361	71.428
	wP	19	5	32	1	56	33.928	8.928	57.142
	NonP	635	257	700	0	1 592	39.886	16.143	43.969

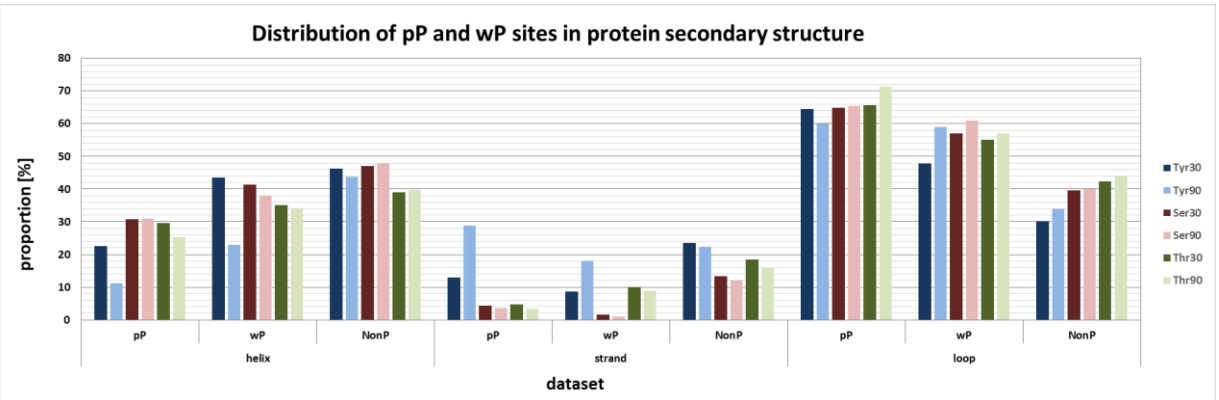


Figure 17: Distribution of pP-, wP- and NonP-sites in protein secondary structure elements ('helix', 'strand', and 'loop' by DSSP (Kabsch and Sander, 1983)). For each amino acid (serine = 'Ser', threonine = 'Thr', and tyrosine = 'Tyr') two sets were made, where the number (30 and 90) corresponds to the maximum sequence identity shared by the protein chains in the redundancy set.

Table 15: P-values of chi-squared analyses of the distributions of pP-, wP-, and NonP-sites in three protein secondary structure elements (helix, sheet, and loop). Chi-squared test was used to find whether the differences between the pP- and wP-sites distributions, pP- and NonP-sites distributions, and wP- and NonP-sites distributions were significant. When the difference between distributions was significant at the 0.05 level, p-value was marked in bold. For each amino acid (serine = 'Ser', threonine = 'Thr', and tyrosine = 'Tyr') two sets were made, where the number (30 and 90) corresponds to the maximum sequence identity shared by the protein chains in the redundancy set.

	pP-wP	pP-NonP	wP-NonP
Tyr30	0.2615	**4.418×10^{-4}**	0.1143
Tyr90	0.0836	**5.366×10^{-9}**	**2.707×10^{-4}**
Ser30	0.2851	**2.410×10^{-7}**	**3.106×10^{-3}**
Ser90	0.2862	**3.964×10^{-10}**	**3.435×10^{-5}**
Thr30	0.4269	**5.167×10^{-4}**	0.2136
Thr90	0.1059	**1.384×10^{-8}**	0.1126

Analysis revealed that non-phosphorylated residues (NonP-sites) had no strict preferences to reside within specific protein secondary structure elements. NonP-sites were found almost equally within alfa helices and loops, and a proportion of NonP-sites within beta sheets was higher than of pP- or wP-sites. On the other hand, pP- as well as wP-sites were both found less in helices and beta sheets and more in loops than NonP-sites. These findings were significant for all pP-sites datasets (the highest p-value was 5.167×10^{-4}) and for wP –sites datasets Tyr90, Ser30 and Ser90 (2.707×10^{-4}, 3.106×10^{-3}, and $3,435 \times 10^{-5}$, respectively). Both pP- and wP-sites distributions in protein secondary structure showed no significant differences (the lowest number of P-value was 0.0836).

5.3 Compactness

Compactness of the site within protein structure can describe the level of folding of protein. Changes in compactness can indicate the strength of interactions between the site and its spatially neighboring amino acids. Average distance decreases when amino acids are moved apart and increases when they begin to be close to each other. This average distance was measured across all datasets described in 5.1. to study whether the pP- sites are compact than wP- and NonP-sites. Amino acids spatially neighboring the site were selected for the analysis by two approaches – Euclidean distance and Voronoi diagrams.

The average distances for all datasets are presented in Supplementary material, table 3 and Figure 18.

Whether these findings were significant was measured using analysis of variance (ANOVA). Zero hypothesis was (H_0):

1. The means of pP- and wP-sites, pP- and NonP-sites, and wP- and NonP-sites average distances are equal.

Results are shown in Supplementary, table 4 and summarized in table 16.

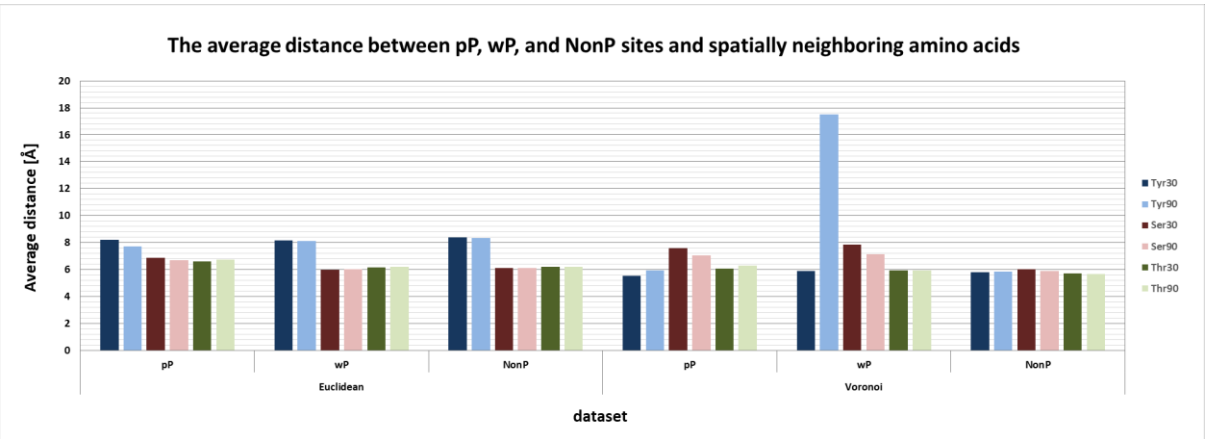


Figure 18: The average distances between the pP-, wP- and NonP-sites and neighboring aa's for both Euclidean distance and Voronoi diagrams. For each amino acid (serine = 'Ser', threonine = 'Thr', and tyrosine = 'Tyr') two sets (30 and 90) were made, where the number corresponds to the maximum sequence identity shared by the protein chains in the redundancy set.

Table 16: P-values of results obtained from ANOVA analyses of differences in sites compactness. For each amino acid (serine = 'Ser', threonine = 'Thr', and tyrosine = 'Tyr') two sets (30 and 90) were made, where the number corresponds to the maximum sequence identity shared by the protein chains in the redundancy set. All three types of sites were compared (pP- with wP-sites, pP- with NonP-sites, and wP- with NonP-sites) for both Euclidean distance and Voronoi diagrams. When the mean difference was significant at the 0.05 level, p-value was marked in bold.

	Euclidean			Voronoi		
dataset	**pP-wP**	**pP-NonP**	**wP-NonP**	**pP-wP**	**pP-NonP**	**wP-NonP**
Tyr30	1.000	0.999	0.999	1.000	1.000	1.000
Tyr90	0.843	**6.892×10^{-4}**	0.999	**$< 2.2 \times 10^{-16}$**	1.000	**$< 2.2 \times 10^{-16}$**
Ser30	**1.105×10^{-3}**	**1.443×10^{-7}**	0.999	1.000	0.099	0.251
Ser90	**9.334×10^{-3}**	**1.263×10^{-6}**	0.999	1.000	0.252	0.633
Thr30	0.960	0.590	0.999	1.000	1.000	1.000
Thr90	0.432	**8.239×10^{-4}**	1.000	1.000	0.999	1.000

Both Euclidean distance and Voronoi diagram showed similar results, except for Tyr90-wP dataset. Whereas aa's spatially neighboring the pP- and wP-sites formed significantly less compact environment than aa's spatially neighboring the NonP-sites for datasets created by Euclidean distance, datasets created by Voronoi diagrams showed no significances except for Tyr90 datasets. Significant differences between the compactness of the sites and spatially neighboring aa's selected by Euclidean distance were found between pP- and wP-sites for serine (1.105×10^{-3} and 9.334×10^{-3} for Ser30 and Ser90, respectively), and between pP- and NonP-sites for Tyr90, Ser30, Ser90, and Thr90 (6.892×10^{-4}, 1.443×10^{-7}, 1.263×10^{-6}, and 8.239×10^{-4}, respectively). To sum up, pP- and wP-sites of Ser and Thr datasets showed lower compactness than NonP-sites, contrary to Tyr pP- and wP-sites.

5.3 Solvent accessibility

Whether pP-, wP-, and NonP-sites tend to be located on the protein surface and exposed to the solvent or buried within protein structure was studied. Relative solvent accessibility of sites of all datasets described in 5.1. are shown in table 17 and Fig. 19.

Chi-squared test was used to find the significant differences between the distributions of pP- and wP-sites (Supplementary material, table 5 (A)), pP- and NonP-sites (Supplementary material, table 5 (B)), and wP- and NonP-sites (Supplementary material, table 5 (C)) to be located on the protein surface (exposed to the solvent), or to be buried within protein structure. Zero hypothesis (H_0) were:

1. Distribution of pP- and wP-sites on the surface and within protein structure is random.
2. Distribution of pP- and NonP-sites on the surface and within protein structure is random.
3. Distribution of wP- and NonP-sites on the surface and within protein structure is random.

Results of chi-squared test analyses are summarized in table 17.

Table 17: The number of exposed and buried pP-, wP- and NonP-sites. For each amino acid two sets (30 and 90) were made, where the number corresponds to the maximum sequence identity shared by the protein chains in the redundancy set. The distributions of exposed and buried sites are presented in numbers ('count') and in relative percentage ('percent'), where sum of all sites in dataset is 100 %.

		count			percent	
dataset	type	exposed	buried	sum	exposed	buried
Tyr30	pP	16	16	32	50.000	50.000
	wP	10	13	23	43.478	56.521
	NonP	107	316	423	25.295	74.704
Tyr90	pP	59	32	91	64.835	35.164
	wP	37	24	61	60.655	39.344
	NonP	248	632	880	28.181	71.818
Ser30	pP	87	30	117	74.358	25.641
	wP	41	22	63	65.079	34.920
	NonP	1 106	713	1 819	60.802	39.197
Ser90	pP	125	41	166	75.301	24.698
	wP	54	38	92	58.695	41.304
	NonP	1 507	978	2 485	60.643	39.356
Thr30	pP	48	16	64	75.000	25.000
	wP	27	13	40	67.500	32.500
	NonP	478	415	893	53.527	46.472
Thr90	pP	75	44	119	63.025	36.974
	wP	34	23	57	59.649	40.350
	NonP	869	723	1 592	54.585	45.414

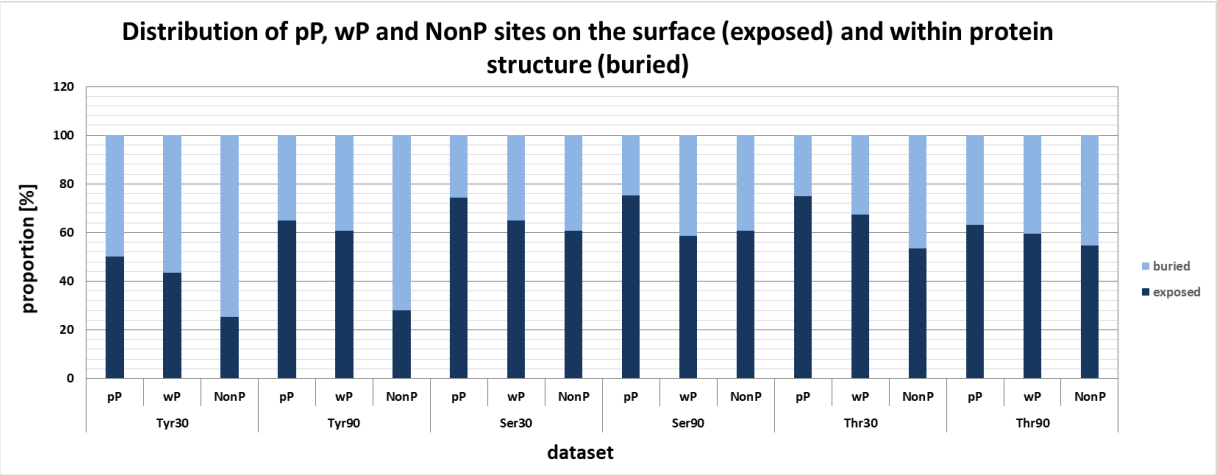


Figure 19: The distributions of exposed and buried pP-, wP- and NonP-sites. For each amino acid (serine = 'Ser', threonine = 'Thr', and tyrosine = 'Tyr') two sets were made, where the number (30 and 90) corresponds to the maximum sequence identity shared by the protein chains in the redundancy set.

Table 17: P-values of chi-squared analyses of the distributions of pP-, wP-, and NonP-sites to be located on protein surface or within protein structure. Chi-squared test was used to find whether the differences between the pP- and wP-sites distributions, pP- and NonP-sites distributions, and wP- and NonP-sites distributions were significant. When the difference between distributions was significant at the 0.05 level, p-value was marked in bold. For each amino acid (serine = 'Ser', threonine = 'Thr', and tyrosine = 'Tyr') two sets were made, where the number (30 and 90) corresponds to the maximum sequence identity shared by the protein chains in the redundancy set.

	pP-wP	pP-NonP	wP-NonP
Tyr30	0.6328	**2.414×10^{-3}**	5.354×10^{-2}
Tyr90	0.6006	**8.147×10^{-13}**	**9.409×10^{-8}**
Ser30	0.1901	**3.47×10^{-3}**	0.4939
Ser90	**5.574×10^{-3}**	**1.708×10^{-4}**	0.7073
Thr30	0.4067	**8.523×10^{-4}**	8.274×10^{-2}
Thr90	0.6660	7.414×10^{-2}	0.4504

Serine and threonine sites were generally more exposed than tyrosine sites. PP- as well as wP-sites were located mainly on the protein surface. Even though the distribution of pP- and wP- sites is similar except Ser90 dataset, phosphosites were significantly located on the protein surface only in phosphorylated state. The percentage of exposed sites was lowest for NonP-sites and highest for pP-sites across all Tyr, Ser, and Thr datasets, except Thr90. The distributions of pP- and NonP- sites on the protein surface or buried within protein structure differed significantly for all datasets except Thr90 (2.414×10^{-3}, 8.147×10^{-13}, 3.47×10^{-3}, 1.708×10^{-4}, and 8.523×10^{-4} for Tyr30, Tyr90, Ser30, Ser90, and Tyr30, respectively). WP- and NonP-sites differed significantly only for Tyr90 (9.409×10^{-8}). The proportion of pP- and wP-sites on the protein surface and within protein structure were similar. Only Ser90 dataset showed significant difference (p-value = 5.574×10^{-3}).

5.4 Hydrophobicity

Hydrophobicity of the amino acids spatially neighboring the site can indicate whether the site is located on the protein surface or within protein structure (buried). It can also give an additional information about the compactness of the site. Whether the proportion of hydrophobic, hydrophilic and polar aa's spatially neighboring the sites is equal for pP-, wP-, and NonP-sites was measured by modified hydrophobic scale, described in 4.2.2.4.

Results are presented in Supplementary material, table 6. Distributions of hydrophobic, hydrophilic and polar aa's in neighborhood of sites are shown in Fig. 20, for aa's selected by Euclidean distance (Fig. 20 (A)), and Voronoi diagrams (Fig. 20 (B)).

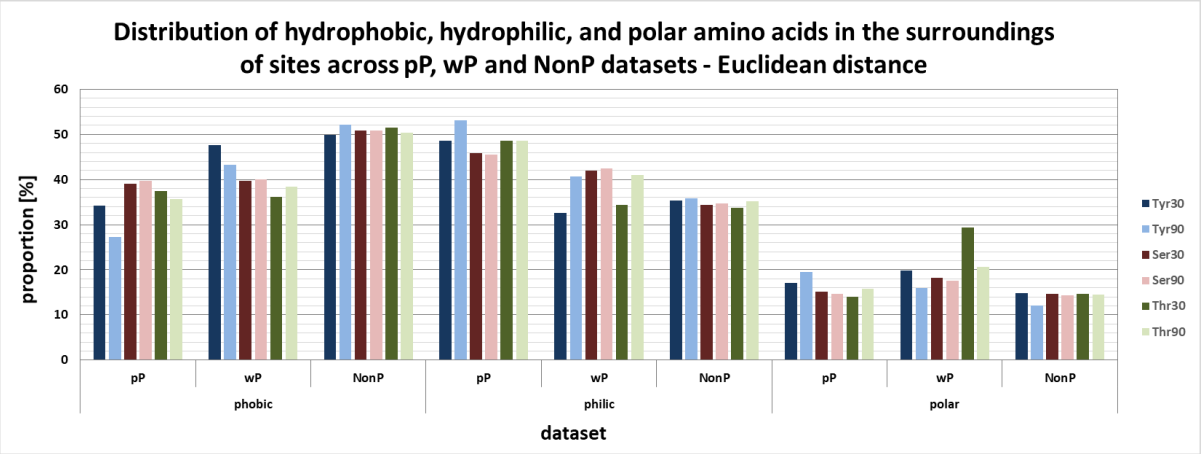
Chi-squared test was used to find the significant differences between the distributions of aa's spatially neighboring the pP- and wP-sites (Supplementary material, table 7 (A)), pP- and NonP-sites (Supplementary material, table 7 (B)), and wP- and NonP-sites (Supplementary material, table 7 (C)) in three categories (hydrophobic, hydrophilic, and polar) based on their hydrophobicity. Chi-squared test was separately used for neighboring aa's selected by Euclidean distance and for neighboring aa's selected by Voronoi diagrams.

Zero hypothesis were:

1. Distribution of aa's spatially neighboring the pP- and wP-sites and selected by Euclidean distance in three categories ('hydrophobic', 'hydrophilic', and 'polar') is random.
2. Distribution of aa's spatially neighboring the pP- and NonP-sites and selected by Euclidean distance in three categories ('hydrophobic', 'hydrophilic', and 'polar') is random.
3. Distribution of aa's spatially neighboring the wP- and NonP-sites and selected by Euclidean distance in three categories ('hydrophobic', 'hydrophilic', and 'polar') is random.
4. Distribution of aa's spatially neighboring the pP- and wP-sites and selected by Voronoi diagrams in three categories ('hydrophobic', 'hydrophilic', and 'polar') is random.
5. Distribution of aa's spatially neighboring the pP- and NonP-sites and selected by Voronoi diagrams in three categories ('hydrophobic', 'hydrophilic', and 'polar') is random.
6. Distribution of aa's spatially neighboring the wP- and NonP-sites and selected by Voronoi diagrams in three categories ('hydrophobic', 'hydrophilic', and 'polar') is random.

Results are summarized in table 18.

(A)



(B)

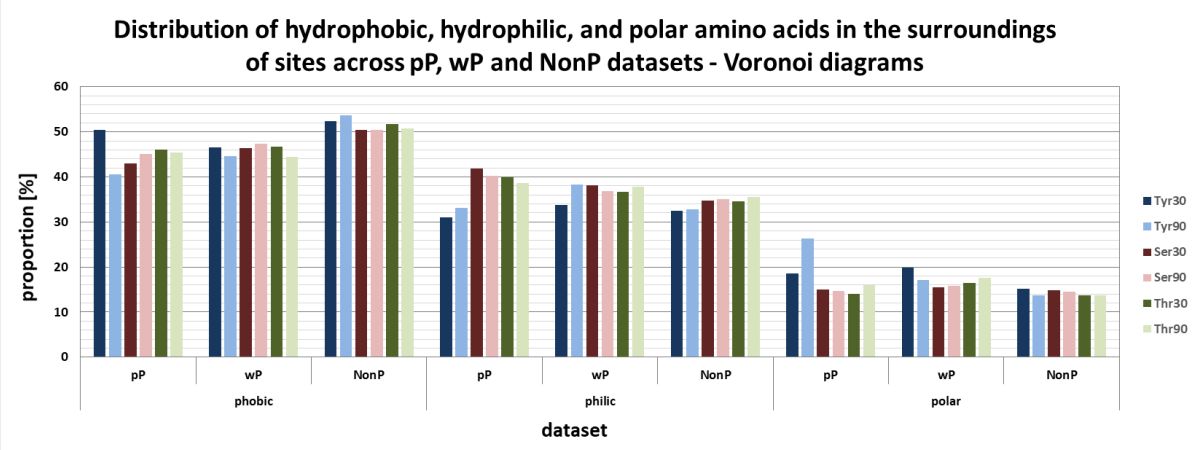


Figure 20: The distribution of aa's spatially neighboring pP-, wP- and NonP-sites in three categories (hydrophobic = 'phobic', hydrophilic = 'philic', and 'polar') based on their hydrophobicity. For each amino acid two sets (30 and 90) were made, where the number corresponds to the maximum sequence identity shared by the protein chains in the redundancy set.

Table 18: P-values of chi-squared analyses of the distributions of aa's spatially neighboring the pP-, wP-, and NonP-sites in three categories ('hydrophobic', 'hydrophilic' and 'polar') based on their hydrophobicity. Chi-squared test was used to find whether the differences between the pP- and wP-sites distributions, pP- and NonP-sites distributions, and wP- and NonP-sites distributions were significant. When the difference between distributions was significant at the 0.05 level, p-value was marked in bold. For each amino acid (serine = 'Ser', threonine = 'Thr', and tyrosine = 'Tyr') two sets were made, where the number (30 and 90) corresponds to the maximum sequence identity shared by the protein chains in the redundancy set.

	Euclidean			Voronoi		
dataset	**pP-wP**	**pP-NonP**	**wP-NonP**	**pP-wP**	**pP-NonP**	**wP-NonP**
Tyr30	**5.125×10^{-3}**	**2.350×10^{-5}**	0.1941	0.7320	0.4264	0.1644
Tyr90	**2.767×10^{-6}**	**$< 2.2 \times 10^{-16}$**	**1.359×10^{-3}**	**1.411×10^{-3}**	**$< 2.2 \times 10^{-16}$**	**8.107×10^{-4}**
Ser30	0.1999	**1.403×10^{-13}**	**3.909×10^{-6}**	0.4506	**2.817×10^{-4}**	0.2763
Ser90	0.1663	**$< 2.2 \times 10^{-16}$**	**3.405×10^{-8}**	0.3981	**2.297×10^{-3}**	0.3123
Thr30	**3.343×10^{-8}**	**1.020×10^{-11}**	**1.013×10^{-14}**	0.6058	8.670×10^{-2}	0.2346
Thr90	**1.430×10^{-2}**	**$< 2.2 \times 10^{-16}$**	**1.010×10^{-6}**	0.7936	**2.053×10^{-2}**	**2.998×10^{-2}**

The percentage of hydrophobic aa's spatially neighboring the sites was higher of NonP-sites than of pP- and wP-sites. Coincidently the percentage of hydrophilic aa's was lower of NonP-sites than of pP- and wP-sites. Polar amino acids were distributed equally between the pP-, wP-, and NonP-sites. These findings were more significant for aa's selected by Euclidean distance than for aa's selected by Voronoi diagrams. For example, distributions of pP- and NonP-sites differed significantly in all datasets created by Euclidean distance for Tyr30, Tyr90, Ser30, Ser90, Thr30, and Thr90 (2.350×10^{-5}, $< 2.2 \times 10^{-16}$, 1.403×10^{-13}, $< 2.2 \times 10^{-16}$, 1.020×10^{-11}, and $< 2.2 \times 10^{-16}$, respectively), but only for Tyr90, Ser30, Ser90, and Thr90 created by Voronoi diagrams ($< 2.2 \times 10^{-16}$, 2.817×10^{-4}, 2.297×10^{-3}, and 2.053×10^{-2}, respectively). Therefore, the distributions of hydrophobic, hydrophilic and polar amino acids depend on the method used for the selection.

Nevertheless, both approaches showed significant differences between the distributions of aa's spatially neighboring the pP-sites and NonP-sites. These findings suggest not only that aa's selected by Euclidean distance can more distinguish between pP-, wP- and NonP-sites, but also that pP-sites can be used in protein phosphosites predictions based on charge of aa's in sites environment.

5.5 Charge

Amino acid spatially neighboring the site bears the positive, negative or no charge according to the functional group of its side chain. Positive charge of the protein site may facilitate the binding of phosphate group on the side chain of serine, threonine, or tyrosine and may stabilize the site by the compensation of the negatively charged phosphate group. Whether pP-, wP-, and NonP-sites are surrounded by the same portion of aa's with positive, negative, or no charge, was studied.

Results are presented in Supplementary material, table 8 and depicted in Fig. 21. Amino acids spatially neighboring the sites were selected by Euclidean distance (Fig. 21 (A)) and Voronoi diagrams (Fig. 21 (B)).

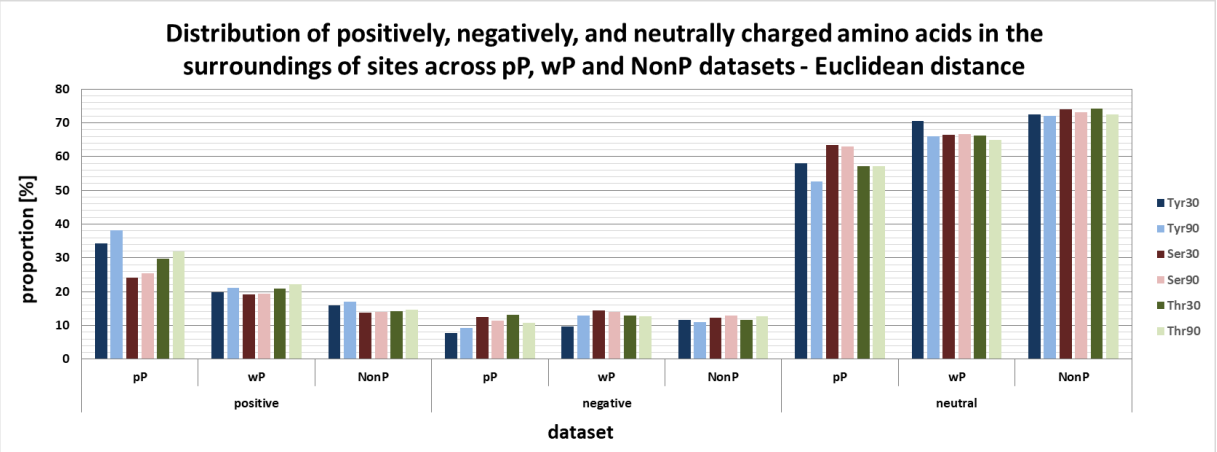
Chi-squared test was used to find the significant differences between the distributions of aa's spatially neighboring the pP- and wP-sites in three categories ('positive', 'negative', and 'neutral') based on their charge. Zero hypothesis (H_0) were:

1. Distribution of aa's spatially neighboring the pP- and wP-sites and selected by Euclidean distance in three categories ('positive, 'negative, and 'neutral) is random.
2. Distribution of aa's spatially neighboring the pP- and NonP-sites and selected by Euclidean distance in three categories ('positive, 'negative, and 'neutral) is random.
3. Distribution of aa's spatially neighboring the wP- and NonP-sites and selected by Euclidean distance in three categories ('positive, 'negative, and 'neutral) is random.
4. Distribution of aa's spatially neighboring the pP- and wP-sites and selected by Voronoi diagrams in three categories ('positive, 'negative, and 'neutral) is random.

5. Distribution of aa's spatially neighboring the pP- and NonP-sites and selected by Voronoi diagrams in three categories ('positive, 'negative, and 'neutral) is random.
6. Distribution of aa's spatially neighboring the wP- and NonP-sites and selected by Voronoi diagrams in three categories ('positive, 'negative, and 'neutral) is random.

Results of chi-squared test are presented in Supplementary material, table 9 and summarized in table 19.

(A)



(B)

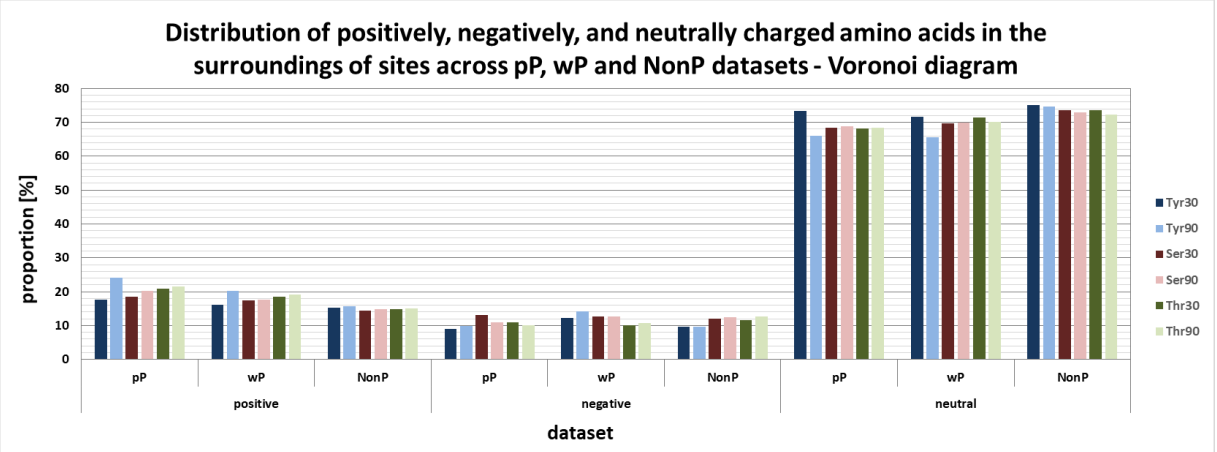


Figure 21: The distribution of positively, negatively, and neutrally charged amino acids spatially neighboring sites across the pP, wP and NonP datasets. Three amino acids as a central residue of the protein site were used – serine (Ser), threonine (Thr), and tyrosine (Tyr). For each amino acid two sets (30 and 90) were made, where the number corresponds to the maximum sequence identity shared by the protein chains in the redundancy set. Method used for the selection of spatially neighboring amino acids was (A) Euclidean distance and (B) Voronoi diagrams.

Table 19: P-values of chi-squared analyses of the distributions of aa's spatially neighboring the pP-, wP-, and NonP-sites. Chi-squared test was used to find significant differences between the distributions of pP- and wP-sites neighboring aa's, pP- and NonP-sites neighboring aa's, and wP- and NonP-sites neighboring aa's. For each amino acid (serine = 'Ser', threonine = 'Thr', and tyrosine = 'Tyr') two sets were made, where the number (30 and 90) corresponds to the maximum sequence identity shared by the protein chains in the redundancy set. When the difference between distributions was significant at the 0.05 level, p-value was marked in bold

	Euclidean			Voronoi		
dataset	**pP-wP**	**pP-NonP**	**wP-NonP**	**pP-wP**	**pP-NonP**	**wP-NonP**
Tyr30	**7.834×10^{-3}**	**1.597×10^{-11}**	0.3434	0.5564	0.6316	0.4396
Tyr90	**1.774×10^{-7}**	**$< 2.2 \times 10^{-16}$**	**2.790×10^{-2}**	0.0513	**2.872×10^{-7}**	**8.392×10^{-5}**
Ser30	7.789×10^{-2}	**$< 2.2 \times 10^{-16}$**	**3.796×10^{-4}**	0.8841	**6.607×10^{-3}**	0.1888
Ser90	**3.911×10^{-3}**	**$< 2.2 \times 10^{-16}$**	**5.498×10^{-5}**	0.3066	**1.595×10^{-5}**	0.1572
Thr30	**1.423×10^{-2}**	**$< 2.2 \times 10^{-16}$**	**1.718×10^{-3}**	0.7126	**9.200×10^{-3}**	0.2330
Thr90	**8.332×10^{-4}**	**$< 2.2 \times 10^{-16}$**	**4.888×10^{-5}**	0.6212	**1.409×10^{-5}**	9.422×10^{-2}

Positively charged amino acids were located more likely in the surroundings of pP-sites than in the surroundings of wP-, or NonP-sites. Furthermore, neutrally charged amino acids were observed mainly in the surroundings of NonP-sites. Amino acids neighboring pP-sites selected by Voronoi diagrams differed significantly from NonP-sites for Tyr90, Ser30, Ser90, Thr30, and Thr90 (p-values 2.872×10^{-7}, 6.607×10^{-3}, 1.595×10^{-5}, 9.200×10^{-3}, and 1.409×10^{-5}, respectively). Distributions of aa's neighboring wP-sites selected by Voronoi diagrams differed significantly from NonP-sites only for Tyr90 (8.392×10^{-5}).

Whereas distributions of aa's spatially neighboring the pP-sites selected by Euclidean distance differed significantly from wP-sites in all datasets except Ser30 (7.834×10^{-3} for Tyr30, 1.774×10^{-7} for Tyr90, 3.911×10^{-3} for Ser90, and 1.423×10^{-2} for Thr30, and 8.332×10^{-4} for Thr90), aa's selected by Voronoi diagrams did not expressed same significant difference. Therefore, the distributions of positively and negatively charged aa's and aa's with no charge depend on the method used for the selection. Nevertheless, distributions of aa's neighboring pP- and wP-sites selected by Euclidean distance differed significantly in all datasets except Tyr30-wP (p-values 1.597 x 10-11 for Tyr30-pP, < 2.2 x 10-16 for Tyr90-pP, Ser30-pP, Ser90-pP, Thr30-pP, and Thr90-pP). These findings suggest not only that aa's selected by Euclidean distance can better distinguish between pP-, wP- and NonP-sites, but also that pP-sites can be used in protein phosphosites predictions based on charge of aa's in sites environment.

5.6 Evolutionary conservation profiles

Evolutionary conservation profiles of phosphosites and aa's neighboring the phosphosites can give an information about the stability of the site and its environment in evolution and thus the importance of this site and aa's spatially neighboring these sites for a cell.

The number of sites used in evolutionary conservation profiles analysis is shown in table 20. ConSurf-DB was not able to give an information about conservation in evolution for all protein chains, not only

because it required at least 50 homologs, but also because several internal errors. Therefore, the number of used protein structures and sites was lower than in previous analyses. Tyrosine pP-set contained 9 and 31 protein structures including 11 and 41 sites for Tyr30 and Tyr90, respectively. Serine pP-sets contained 29 and 44 protein structures including 36 and 61 sites for Ser30 and Ser90, respectively. Threonine pP-sets contained 10 and 33 protein structures including 12 and 37 sites for Thr30 and Thr90, respectively. Tyrosine wP-set contained 8 and 20 protein structures including 8 and 27 sites for Tyr30 and Tyr90, respectively. Serine wP-sets contained 19 and 28 protein structures including 25 and 35 sites for Ser30 and Ser90, respectively. Threonine wP-sets contained 9 and 14 protein structures including 14 and 20 central residues for Thr30 and Thr90, respectively. Tyrosine NonP-set contained 9 and 31 protein structures including 174 and 409 sites for Tyr30 and Tyr90, respectively. Serine NonP-sets contained 29 and 44 protein structures including 400 and 734 sites for Ser30 and Ser90, respectively. Threonine NonP-sets contained 10 and 33 protein structures including 82 and 422 sites for Thr30 and Thr90, respectively.

Table 20: The number of protein structures and sites used in evolutionary conservation profiles analysis of pP-, wP-, and NonP-sites. For each amino acid (serine = 'Ser', threonine = 'Thr', and tyrosine = 'Tyr') two sets were made, where the number (30 and 90) corresponds to the maximum sequence identity shared by the protein chains in the redundancy set.

	pP-sets		wP-sets		NonP-sets	
	Protein structures	sites	Protein structures	sites	Protein structures	sites
Tyr30	9	11	8	8	9	174
Tyr90	31	41	20	27	31	409
Ser30	29	36	19	25	29	400
Ser90	44	61	28	35	44	734
Thr30	10	12	9	14	10	82
Thr90	33	37	14	20	33	422

The average number of conservations of the sites and their neighboring aa's are presented in Supplementary material, table 10 and depicted in Fig. 22 and Fig. 23. Average conservation is expressed in a range (1-9), where 1 is the most variable one, 9 is the most conserved one.

Analysis of variance (ANOVA) was used to find significant differences between datasets. Zero hypothesis (H_0) were:

1. The means of pP- and wP-sites, pP- and NonP-sites, and wP- and NonP-sites conservation levels are equal.
2. The means of aa's spatially neighboring the pP- and wP-sites, pP- and NonP-sites, and wP- and NonP-sites conservation levels are equal.

Results are shown in Supplementary material, table 11 and summarized in table 21 and table 22.

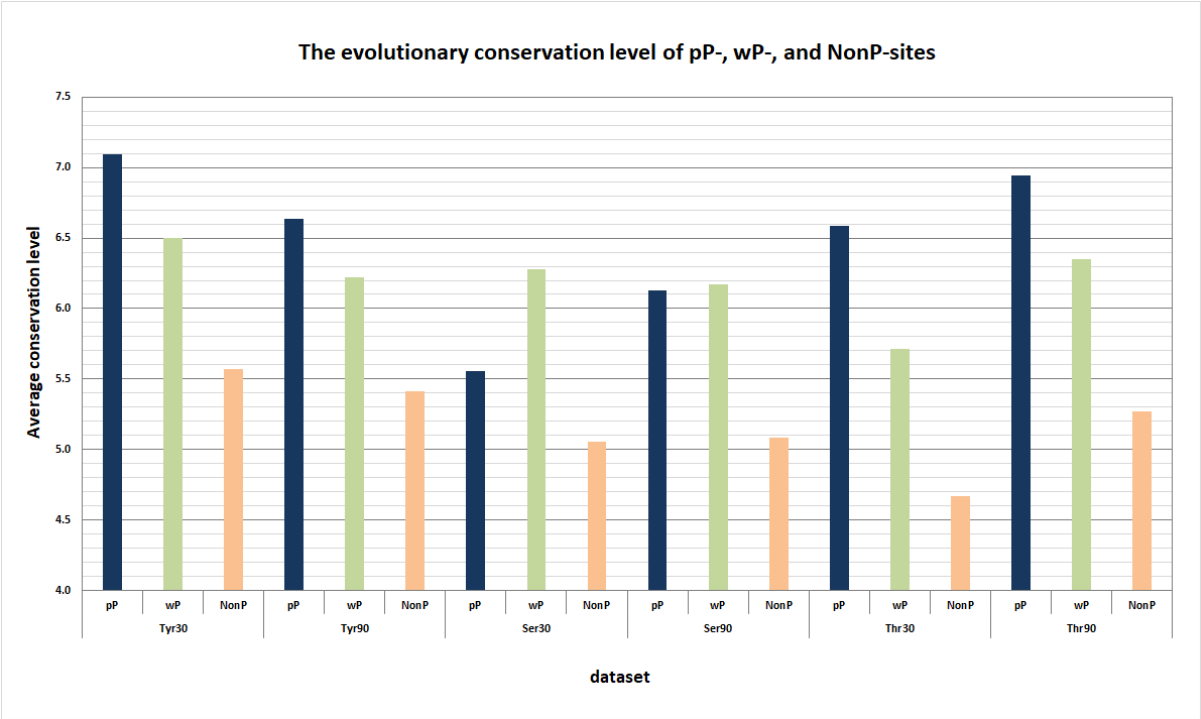


Figure 22: The evolutionary conservation profiles of pP-, wP- and NonP-sites. For each amino acid (serine = 'Ser', threonine = 'Thr', and tyrosine = 'Tyr') two sets were made, where the number (30 and 90) corresponds to the maximum sequence identity shared by the protein chains in the redundancy set.

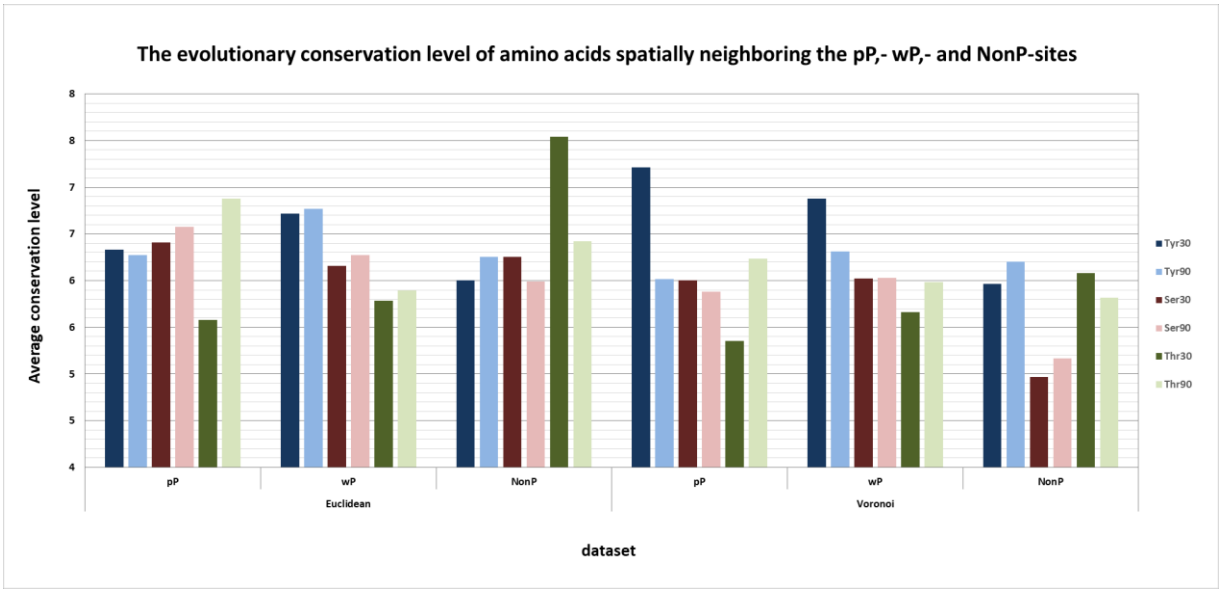


Figure 23: The evolutionary conservation profiles of aa's spatially neighboring the pP-, wP- and NonP-sites. For each amino acid (serine = 'Ser', threonine = 'Thr', and tyrosine = 'Tyr') two sets were made, where the number (30 and 90) corresponds to the maximum sequence identity shared by the protein chains in the redundancy set. Amino acids were selected by Euclidean distance ('Euclidean', on the left), and Voronoi diagrams ('Voronoi', on the right).

Table 21: P-values of ANOVA analyses of the evolutionary conservation profiles of pP-, wP-, and NonP-sites. ANOVA was used to find significant differences between pP- and wP-sites, between pP- and NonP-sites, and between wP- and NonP-sites. For each amino acid (serine = 'Ser', threonine = 'Thr', and tyrosine = 'Tyr') two sets were made, where the number (30 and 90) corresponds to the maximum sequence identity shared by the protein chains in the redundancy set. When the mean difference was significant at the 0.05 level, p-value was marked in bold.

	pP-wP	pP-NonP	wP-NonP
Tyr30	1.000	0.930	1.000
Tyr90	1.000	0.308	0.988
Ser30	1.000	1.000	0.721
Ser90	1.000	0.634	0.222
Thr30	1.000	0.656	0.997
Thr90	1.000	**0.026**	0.950

Table 22: P-values of ANOVA analyses of the evolutionary conservation profiles of aa's spatially neighboring the pP-, wP-, and NonP-sites. ANOVA was used to find significant differences between aa's neighboring the pP- and wP-sites, pP- and NonP-sites, and wP- and NonP-sites. For each amino acid (serine = 'Ser', threonine = 'Thr', and tyrosine = 'Tyr') two sets were made, where the number (30 and 90) corresponds to the maximum sequence identity shared by the protein chains in the redundancy set. Amino acids were selected by Euclidean distance 'Euclidean') or Voronoi diagrams ('Voronoi'). When the mean difference was significant at the 0.05 level, p-value was marked in bold.

	Euclidean			Voronoi		
dataset	**pP-wP**	**pP-NonP**	**wP-NonP**	**pP-wP**	**pP-NonP**	**wP-NonP**
Tyr30	1.000	1.000	0.999	1.000	**0.043**	0.659
Tyr90	0.997	1.000	0.970	1.000	1.000	1.000
Ser30	1.000	1.000	1.000	1.000	**$< 2.2 \times 10^{-16}$**	**$< 2.2 \times 10^{-16}$**
Ser90	1.000	0.473	1.000	1.000	**0.001**	**0.001**
Thr30	1.000	**0.006**	**0.007**	1.000	0.822	0.998
Thr90	0.635	0.952	0.991	1.000	0.708	1.000

Even though at first sight pP- and wP- sites seem to be more conserved than NonP-sites, analysis revealed, that no differences between pP- and wP-sites, pP- and NonP-sites, as well as wP- and NonP-sites were significant except between pP- and NonP-sites of Thr90 (p-value 0.026). Amino acids spatially neighboring the pP- and wP-sites were significantly more conserved than aa's neighboring the NonP-sites in Ser30 and Ser90 datasets, when they were selected by Voronoi diagrams (p-value $< 2.2 \times 10^{-16}$ for Ser30-pP and Ser90-wP, and 0.001 for Ser90-pP and Ser90-wP). Amino acids selected by Euclidean distance were significantly more conserved around the pP- and wP- sites than around the

NonP-sites in Thr30 datasets (0.006 for pP sites, 0.007 for wP-sites). Therefore, contrary to presumption, pP- and wP-sites and aa's spatially neighboring them were not more conserved than NonP-sites.

5.7 Conformational change upon phosphorylation

The position of the phosphosite side chain was identified before and after phosphorylation event using pairs of protein chains with the same UniProt accession ID that were superposed. RMSD was then measured across all datasets described in 5.1. Because change in position of amino acid's atoms bigger equals or bigger than 3.5 Å suggest a great protein conformational change, two datasets were made, one with samples having RMSD < 3.5 Å, second with all samples. The number of sites and its average RMSD is presented in the following table 23.

Table 23: The number of sites used in RMSD analysis of pP-, wP-, and NonP-sites and datasets average RMSD values. For each amino acid (serine = 'Ser', threonine = 'Thr', and tyrosine = 'Tyr') two sets were made, where the number (30 and 90) corresponds to the maximum sequence identity shared by the protein chains in the redundancy set. Two datasets were created, one contains all samples, second one with samples having RMSD less than 3.5 Å.

	all samples		samples with RMSD < 3.5 Å	
dataset	**number**	**average [Å]**	**number**	**Average [Å]**
Tyr30	18	6.1992	9	1.3475
Tyr90	55	9.3502	26	1.6147
Ser30	53	7.6239	34	1.2477
Ser90	73	4.5272	54	1.2616
Thr30	27	4.3128	18	1.5220
Thr90	41	6.3965	27	1.3085

Whether the change in position of amino acid's atoms was significant, was analyzed by one sample t-test. Zero hypothesis (H_0) was:

1. Average differences of amino acid atom positions before and after the phosphorylation equals zero.

Results of analysis are presented in table 24.

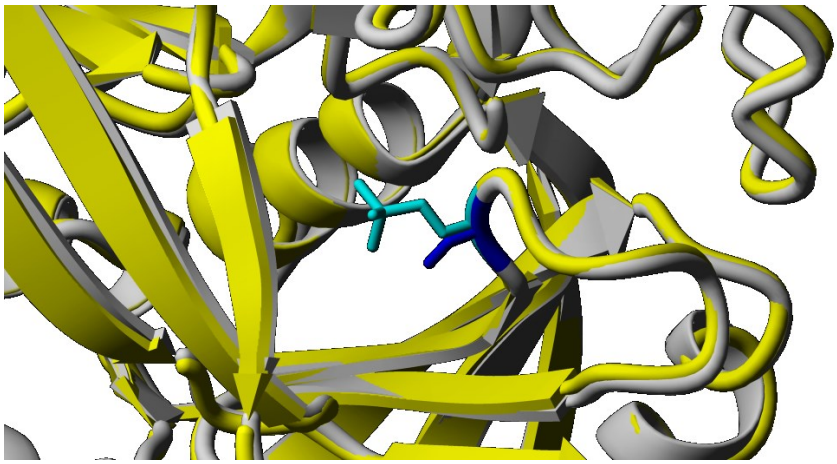
Table 24: P-values of one sample t-test analyses of the atoms position change of pP-, wP-, and NonP-sites. For each amino acid (serine = 'Ser', threonine = 'Thr', and tyrosine = 'Tyr') two sets were made, where the number (30 and 90) corresponds to the maximum sequence identity shared by the protein chains in the redundancy set. Two datasets were further created, one contains all samples and second one contains only samples having RMSD less than 3.5 Å. T-value, which represents a size of the difference relative to the variation in sample data, degrees of freedom ('df' - the number of parameters of the system that may vary independently), and p-value is presented. When the phosphosite side chain position in space before and after phosphorylation was different significantly at the 0.05 level, p-value was marked in bold.

	all samples			samples with RMSD < 3.5 Å		
dataset	**t-value**	**df**	**p-value**	**t-value**	**df**	**p-value**
Tyr30	2.4022	17	**0.028**	10.9560	8	**4.276×10^{-6}**
Tyr90	6.5621	54	**2.115×10^{-8}**	11.1820	25	**3.210×10^{-11}**
Ser30	4.3196	52	**7.057×10^{-5}**	10.6090	33	**3.612×10^{-12}**
Ser90	4.9782	72	**4.239×10^{-6}**	13.4490	53	**$< 2.2 \times 10^{-16}$**
Thr30	4.7822	26	**5.971×10^{-5}**	8.1468	17	**2.847×10^{-7}**
Thr90	4.5551	40	**4.829×10^{-5}**	10.1250	26	**1.635×10^{-10}**

A number of proteins went through large conformational change after the phosphorylation (9 from 19 of Tyr30, 29 from 55 of Tyr90, 19 from 53 of Ser30, 19 from 73 of Ser90, 9 from 27 of Thr30, and 14 from 41 of Thr90). Tyrosines are more likely to undergo a large conformational change upon phosphorylation. Datasets of samples having RMSD values less than 3.5 Å had smaller variation and, therefore, t-values was higher than in all samples. Further, smaller dataset showed higher significances of changes, because t-test worked with more normal distribution than in dataset with all samples (having small amount of distinct values). Nevertheless, both datasets showed significant difference in amino acid position in space before and after the phosphorylation. Side chains of phosphosites moved after phosphorylation significantly in datasets including all samples (0.028, 2.115×10^{-8}, 7.057×10^{-5}, 4.239×10^{-6}, 5.971×10^{-5}, and 4.829×10^{-5} for Tyr30, Tyr90, Ser30, Ser90, Thr30, and Thr90), as well as in datasets including samples with RMSD less than 3.5 Å (4.276×10^{-6}, 3.210×10^{-11}, 3.612×10^{-12}, $< 2.2 \times 10^{-16}$, 2.847×10^{-7}, and 1.635×10^{-10} for Tyr30, Tyr90, Ser30, Ser90, Thr30, and Thr90). Results were same also for the analysis conducted with RMSD values calculated only for Cβ atoms.

To illustrate a small and a large conformational change of protein structure upon phosphorylation, two cases are depicted: one with RMSD less than 3.5 Å and one with RMSD higher than 3.5Å (Fig. 24). Protein structure containing phosphosite which atoms after phosphorylation changed the position in average 0.8932 Å is shown in Fig. 24 (A). Protein structure containing phosphosite which atoms after phosphorylation changed the position in average 24.9743 Å is shown in Fig. 24 (B).

(A)



(B)

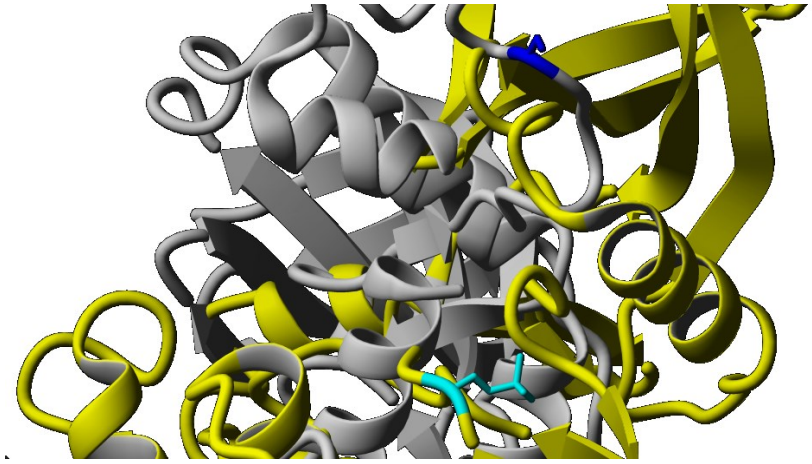


Figure 24: Protein structures containing phosphosites with RMSD (A) smaller than 3.5 Å (B) bigger than 3.5 Å. Two superposed protein structures are presented for each: (A) protein structure containing non-phosphorylated phosphosite on the position 108 (4mrq), protein structure containing phosphorylated phosphosite on the same position (108) (1k35). (B) Protein structure containing non-phosphorylated phosphosite on the position 172 (4eut), protein structure containing phosphorylated phosphosite on the same position (172) (4euu). Protein structure containing phosphorylated phosphosite was colored yellow, containing non-phosphorylated phosphosite gray. Phosphorylated phosphosite was colored light blue, non-phosphorylated phosphosite dark blue. Pictures were made using YASARA tool (Krieger and Vriend, 2014).

6. Discussion

Recent analyses of phosphorylation sites studied mainly non-phosphorylated phosphorylation sites and the distribution and representation of amino acids sequentially neighboring them (Chou, 2016). However, three-dimensional properties of phosphosites may play an important role in the recognition of substrates by protein kinases. Computational methods using three-dimensional properties of phosphosites were proposed to achieve a better prediction accuracy (Durek et al., 2009). Even if the presumption that phosphosites reside preferentially in intrinsically disordered regions of proteins proves to be true (Tyanova et al., 2013), (Zilberstein et al., 2011), (Gao et al., 2010), and (Iakoucheva et al., 2004), structure-based properties of phosphosites will be still beneficial because the number of phosphosites in proteomes is generally fairly high (Vlastaridis et al., 2017).

Therefore, the three main aims of this thesis were to study spatial environment of phosphorylated phosphosites, to compare the obtained results with the results of non-phosphorylated phosphosites analyses, and to find out how much phosphorylation changes the conformational structure of phosphosites and even whole proteins.

For these analyses, non-redundant datasets containing phosphosites within protein structures with proper-folded domains, high resolution and appropriate R-values were created. These datasets as well as Python scripts made for this purpose could be used in following analyses, or in the development of a new protein phosphosite prediction tool. Therefore, the data are attached on a CD (Additional supplementary material 1) and protein chains used for the analyses are presented in Additional supplementary material 2.

A list of structures with phosphorylated phosphosites (1031, 869, and 566 protein structures for pSer, pThr, and pTyr respectively) were obtained from the PDBeChem database (Dimitropoulos et al., 2006) from 9. 7. 2018. These structures were downloaded from the PDB. However, only proper-folded domains of sufficient length (not peptides) could be used in analysis and, therefore, protein chains with a length less than 30 aa's were excluded. The same filter has been already used; for example for UniProt database (Bateman et al., 2017). Besides, protein structures with R-values more than 0.3 and with a resolution higher than 3.5 were filtered out to ensure that only high-quality data were included. Redundancy of datasets were removed using CD-HIT (Huang et al., 2010). For each dataset, two subsets (30 and 90) were created, where the number corresponds to the maximum sequence identity shared by the protein chains in the redundancy set. In the end, only about 2.5 %, 7.2 %, 5.6 %, 7.9 %, 4.1 %, and 7.7 % of phosphosites across Tyr30, Tyr90, Ser30, Ser90, Thr30, and Thr90 datasets remained from the original lists of phosphosites.

Non-phosphorylated phosphosites were obtained from UniProt database (Bateman et al., 2017) using the same accession ID as phosphorylated phosphosites had. Then, these datasets went through the same filtering steps as phosphorylated phosphosites (described above). The number of non-phosphorylated

phosphosites followed the same trend. From the original lists of non-phosphorylated phosphosites remained about 0.67 %, 1.77 %, 1.60 %, 2.34 %, 1.14 %, and 1.62 % of phosphosites across Tyr30, Tyr90, Ser30, Ser90, Thr30, and Thr90 datasets, respectively.

One potential source of error in our study is the way we constructed the negative dataset - all serine, threonine and tyrosine residues not-annotated as phosphorylated were considered to be negative sites. Because the resulting negative sites were exposed to phosphorylation reaction conditions, the probability that they include false negative sites/ones is low. However, this approach cannot guarantee that the sites in NonP-sets are not phosphorylated and they might be determined to have been phosphorylated (and thus false negative) in the future. Moreover, it is important to note that the insight into tyrosine phosphosites properties was limited in this study because only few protein structures containing tyrosine phosphosites were structurally solved.

The properties of both phosphorylated and non-phosphorylated phosphosites were analyzed. For the analysis of phosphosites environments amino acids spatially neighboring the phosphosites were used. These amino acids were selected using two independent methods, namely Euclidean distance and Voronoi diagrams. The number of aa's selected by Euclidean distance as neighboring phosphosite differed from the number of aa's selected by Voronoi diagrams. Voronoi diagrams selected fewer aa's as neighbors than Euclidean distance with a small exception in Tyr30 and Tyr90 datasets, where Voronoi diagrams found more neighbors. This is a consequence of the method used in the creation of Voronoi diagrams: because phosphate is attached on an atom that is in the case of tyrosine farther from the peptide backbone than in the case of serine or threonine, Voronoi diagrams gets/has the possibility to select aa's located farther from the phosphosite. The Euclidian distance method allows to identify all potential binding partners of the phosphosite, while Voronoi diagrams can miss some interaction partners because there are other atoms that are closer to the phosphosite. This was reflected in the results by the fact that statistically significant differences were often detected between the neighborhoods identified through Euclidian distance, but not statistically significant for the differences detected between neighborhoods identified through Voronoi diagrams. Therefore, Euclidean distance was considered to be a more relevant method to address the aims.

A secondary structure state preference of the phosphosites (alpha helices, beta strands, or loops) was studied. According to the analysis of annotations obtained from SIFTS files (Velankar et al., 2013) phosphosites were localized preferentially in loops, whereas non-phosphorylated residues (serine, threonine and tyrosine) were almost equally distributed between alpha helices and loops. Furthermore, non-phosphorylated serine, threonine and serine residues were more often found in beta strands than phosphosites. Phosphosites significantly tend to be located in less structured, flexible protein regions such as loops. This finding is in agreement with previous studies ((Karabulut and Frishman, 2016), (Durek et al., 2009), (Jiménez et al., 2007), (Gnad et al., 2007)) that showed the tendency of phosphosites

to reside within loops and hinges. Loops and hinges are supposed to have similar properties as the disordered regions of proteins.

Compactness of phosphosites was studied to establish whether phosphosites tend to reside within more compact environments than non-phosphorylated residues. Compactness of site was defined by the structural position of side chains of aa's spatially neighboring phosphosites and so the average Euclidean distance between these side chains and phosphosites was measured. It was found out that phosphosites occurred preferentially in less compact environments. In addition, sites after phosphorylation were more relaxed – aa's spatially neighboring these sites were moved apart. To the author's knowledge there is no study dealing with compactness of phosphosites. Nevertheless, such a study may prove useful for protein phosphosites predictions in intrinsically disordered proteins, where compactness of sites is beneficial additional information as shown in (Konrat, 2009).

Solvent accessibility of phosphosites was studied to examine the necessity of phosphosites to be accessible to protein kinases. Interestingly, whereas phosphorylated phosphosites were significantly more often located on protein surfaces, non-phosphorylated phosphosites showed less or no significance. These findings are in contradictionto previous results reported in literature (Gnad et al., 2007), (Jiménez et al., 2007), (Frades et al., 2015), where non-phosphorylated phosphosites were suggested to be significantly more frequently located on protein surfaces. These results thus need to be interpreted with caution. The most likely explanation is that the solvent accessibility prediction tools differ in accuracy. Because NetSurfP-2.0 (Klausen et al.) used in this analysis is a new solvent accessibility prediction tool, no unbiased comparison of its accuracy has not been published yet.

Charge and hydrophobicity of aa's spatially neighboring phosphosites were studied/examined to stress how important the contribution of spatially neighboring aa's is. Several recently published papers showed the importance of charge and hydrophobicity of phosphosite sequentially neighboring aa's (Karabulut and Frishman, 2016), (Huang et al., 2015), (Frades et al., 2015)). Fan and Zhang also stated the importance of charge and hydrophobicity of spatially neighboring aa's (Fan and Zhang, 2005). This thesis confirms the previousfindings that aa's spatially neighboring phosphosites are mainly hydrophilic and positively charged.

Conservation of phosphosites was also studied. The level of conservation was different in phosphorylated phosphosites than in non-phosphorylated residues, but no significance was found. These findings are in agreement with (Miao et al., 2018), but in contradiction to other research carried out in this area ((Mann et al., 2007). As hypothesized by (Miao et al., 2018), it may be a result of a high-turnover of newly emerged phosphosites. Given that these findings are based on a limited number of phosphosites and encountered problems with ConSurf-DB, the results from such analyses should thus be treated with considerable caution. In prospective analyses, alignments should be made without

external tools. Besides, tyrosine residues might be analyzed specifically for vertebrate orthologues because of the increased importance of tyrosine phosphorylation for vertebrates (Tan et al., 2009).

A general trend was observed across analyses that the properties of non-phosphorylated phosphosites were less distinguishable from the properties of non-phosphorylated residues compared to phosphorylated phosphosites.

Another important issue was whether phosphorylated phosphosites can be used in phosphosites feature extraction. Preferences for protein secondary structures as well as compactness of sites did not depend on the phosphorylation state of phosphosites. On the other hand, hydrophobicity and charge of aa's spatially surrounding phosphosites differ between phosphorylated and non-phosphorylated phosphosites. Therefore, the analysis of conformational change of phosphosites upon phosphorylation was conducted. This analysis was done by measuring average distance changes of atoms of phosphosites and it revealed that side chains of phosphosites were moved significantly in the space upon phosphorylation. Furthermore, it was found that 37 % of phosphosites went through large conformational change (more than 3.5 Å) upon phosphorylation. Tyrosines were more likely to experience a large conformational change upon phosphorylation. Therefore, in almost half of the cases two different environments were compared in charge and hydrophobicity analyses. It may explain the differences between charge and hydrophobicity of spatially neighboring aa's of phosphorylated and non-phosphorylated phosphosites. This finding suggests a future direction of our research – can we predict which phosphorylation event will lead to a major conformational change?

7. Summary

This thesis outlines the importance of structure-based information in protein phosphosites annotation and showed the possibility of using of phosphosites in phosphorylation state in analyses.

Firstly, recently published papers dealing with structural properties of phosphosite and protein phosphosites prediction tools were briefly introduced. Secondly, the properties of phosphosites and aa's spatially neighboring the phosphosites were investigated to see whether they significantly distinguish phosphosites from the sites where phosphorylation never occur. Six different properties (protein secondary structure, compactness, solvent accessibility, hydrophobicity, charge, evolutionary conservation and the change of phosphosite side chain position in space) were studied.

Several significant properties were found. Loops and hinges harbored a higher proportion of phosphosites and a lower proportion of non-phosphorylated residues. It was supposed to have a relation to the other finding that phosphosites resided generally in less compact environments within a protein structure., Phosphosites were found mostly on protein surfaces (thus solvent accessible) and the aa's spatially neighboring the phosphosites were more often hydrophilic and bearing a positive charge. Phosphosites were not significantly more conserved than non-phosphorylated residues. Finally, the positions of phosphosites' side chains in the space were significantly different before and after phosphorylation. The analysis of conformational change of phosphosites' side chains upon phosphorylation also revealed that phosphorylation causes in 37 % of cases large conformational changes of protein structures.

Structural properties of phosphorylated phosphosites were also studied in this thesis. The results show that phosphorylated phosphosites do not differ in several properties (in protein secondary structure elements preferences, compactness, and solvent accessibility) from non-phosphorylated phosphosites and could be thus used as training data for phosphosite prediction tools.

However, other studied characteristics (hydrophobicity and charge of spatially neighboring aa's) differed between phosphorylated phosphosites and non-phosphorylated phosphosites. It was caused by a large conformational change induced by phosphorylation in about 37 % of cases. In future analyses, these cases should be analyzed to see what determines that phosphorylation leads to important conformational changes of protein structures. Moreover, further studies, which will use carefully manually edited alignments in the analysis of phosphosites conservation, will need to be undertaken.

The acquired datasets and analysis results will be used in a running collaboration with the group of David Hoksza at MFF to develop a phosphosite prediction tool.

8. List of references

Adamczak, R., Porollo, A., and Meller, J. (2005). Combining prediction of secondary structure and solvent accessibility in proteins. Proteins Struct. Funct. Genet. *59*, 467–475.

Akutsu, T., Zhang, Z., Song, J., Wang, H., Yang, B., Wang, J., Leier, A., Marquez-Lago, T., Webb, G.I., and Daly, R.J. (2017). PhosphoPredict: A bioinformatics tool for prediction of human kinase-specific phosphorylation substrates and sites by integrating heterogeneous feature selection. Sci. Rep. *7*, 1–19.

Aurenhammer, F. (1991). Voronoi diagrams---a survey of a fundamental geometric data structure. ACM Comput. Surv. *23*, 345–405.

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. *28*, 45–48.

Barber, C.B., and Dobkin, D.P. (1996). The Quickhull Algorithm for Convex Hulls. ACM Trans. Math. Softw. *22*, 469–483.

Bateman, A., Martin, M.J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., Bely, B., Bingley, M., Bonilla, C., Britto, R., et al. (2017). UniProt: The universal protein knowledgebase. Nucleic Acids Res. *45*, 158–169.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

Berry, E.A., Dalby, A.R., and Yang, Z.R. (2004). Reduced bio basis function neural network for identification of protein phosphorylation sites: Comparison with pattern recognition algorithms. Comput. Biol. Chem. *28*, 75–85.

Blom, N., Kreegipuu, A., and Brunak, S. (1998). PhosphoBase: a database of phosphorylation sites. Nucleic Acids Res. *26*, 384–386.

Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S., and Brunak, S. (2004). Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics *4*, 1633–1649.

Brinkworth, R.I., Breinl, R.A., and Kobe, B. (2003). Structural basis and prediction of substrate specificity in protein serine/threonine kinases. Proc. Natl. Acad. Sci. *100*, 74–79.

Brinkworth, R.I., Huber, T., Kobe, B., Saunders, N.F., and Kemp, B.E. (2008). Predikin and PredikinDB: a computational framework for the prediction of protein kinase peptide specificity and an associated database of phosphorylation sites. BMC Bioinformatics *9*, 1–11.

Brister, J.R., Ako-Adjei, D., Bao, Y., and Blinkova, O. (2015). NCBI viral genomes resource. Nucleic Acids Res. *43*, 571–577.

Cesaro, L., and Pinna, L.A. (2015). The generation of phosphoserine stretches in phosphoproteins: Mechanism and significance. Mol. Biosyst. *11*, 2666–2679.

Chen, S.C., Chen, F., and Li, W. (2010). Phosphorylated and nonphosphorylated serine and threonine residues evolve at different rates in mammals. Mol. Biol. Evol. *27*, 2548–2554.

Chou, Y.X. and K.-C. (2016). Recent Progress in Predicting Posttranslational Modification Sites in Proteins. Curr. Top. Med. Chem. *16*, 591–603.

Cohen, P. (2002). The origins of protein phosphorylation. Nat Cell Biol *4*, 127–130.

Cousins, K.R. (2005). ChemDraw Ultra 9.0. J. Am. Chem. Soc. *127*, 4115–4116.

Coutsias, E.A., Seok, C., and Dill, K.A. (2004). Using quaternions to calculate RMSD. J. Comput. Chem. *25*, 1849–1857.

Creixell, P., Schoof, E.M., Tan, C.S.H., and Linding, R. (2012). Mutational properties of amino acid residues: implications for evolvability of phosphorylatable residues. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *367*, 2584–2593.

Dang, T.H., Van Leemput, K., Verschoren, A., and Laukens, K. (2008). Prediction of kinase-specific phosphorylation sites using conditional random fields. Bioinformatics *24*, 2857–2864.

Diella, F., Cameron, S., Gemünd, C., Linding, R., Via, A., Kuster, B., Sicheritz-Pontén, T., Blom, N., and Gibson, T.J. (2004). Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. BMC Bioinformatics *5*, 1–5.

Dimitropoulos, D., Ionides, J., and Henrick, K. (2006). Using MSDchem to Search the PDB Ligand Dictionary. Curr. Protoc. Bioinforma. *15*, 1–21.

Dinkel, H., Chica, C., Via, A., Gould, C.M., Jensen, L.J., Gibson, T.J., and Diella, F. (2011). Phospho.ELM: A database of phosphorylation sites-update 2011. Nucleic Acids Res. *39*, 261–267.

Dou, Y., Yao, B., and Zhang, C. (2014). PhosphoSVM: Prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine. Amino Acids *46*, 1459–1469.

Duan, G., and Walther, D. (2015). The roles of post-translational modifications in the context of protein interaction networks. PLoS Comput. Biol. *11*, 1–23.

Durek, P., Schudoma, C., Weckwerth, W., Selbig, J., and Walther, D. (2009). Detection and characterization of 3D-signature phosphorylation site motifs and their contribution towards improved phosphorylation site prediction in proteins. BMC Bioinformatics *10*, 1–17.

Edelsbrunner, H., and Seidel, R. (1986). Voronoi diagrams and arrangements. Discrete Comput. Geom. *1*, 25–44.

Fan, S.C., and Zhang, X.G. (2005). Characterizing the microenvironment surrounding phosphorylated protein sites. Genomics, Proteomics Bioinforma. *3*, 213–217.

Frades, I., Resjö, S., and Andreasson, E. (2015). Comparison of phosphorylation patterns across eukaryotes by discriminative N-gram analysis. BMC Bioinformatics *16*, 1–13.

Gao, J., Thelen, J.J., Dunker, A.K., and Xu, D. (2010). Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. Mol. Cell. Proteomics *9*, 2586–2600.

Gao, Y., Hao, W., Gu, J., Liu, D., Fan, C., Chen, Z., and Deng, L. (2016). PredPhos: An ensemble framework for structure-based prediction of phosphorylation sites. J. Biol. Res. *23*, 29–39.

Gnad, F., Ren, S., Cox, J., Olsen, J. V, Macek, B., Oroshi, M., and Mann, M. (2007). PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. Genome Biol *8*, 1–13.

Goldenberg, O., Erez, E., Nimrod, G., and Ben-Tal, N. (2009). The ConSurf-DB: Pre-calculated evolutionary conservation profiles of protein structures. Nucleic Acids Res. *37*, 323–327.

Heazlewood, J.L., Durek, P., Hummel, J., Selbig, J., Weckwerth, W., Walther, D., and Schulze, W.X. (2008). PhosPhAt: A Database of phosphorylation sites in Arabidopsis thaliana and a plant specific phosphorylation site predictor. Nucleic Acids Res *36*, 1015–1021.

Hjerrild, M., and Gammeltoft, S. (2006). Phosphoproteomics toolbox: Computational biology, protein chemistry and mass spectrometry. FEBS Lett. *580*, 4764–4770.

Holm, L., and Laakso, L.M. (2016). Dali server update. Nucleic Acids Res. *44*, 351–355.

Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J.M., Latham, V., and Skrzypek, E. (2015). PhosphoSitePlus, 2014: Mutations, PTMs and recalibrations. Nucleic Acids Res. *43*, 512–520.

Hornbeck, P. V, Kornhauser, J.M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., and Sullivan, M. (2012). PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res. *40*, 261–270.

Hou, Q., Bourgeas, R., Pucci, F., and Rooman, M. (2018). Computational analysis of the amino acid interactions that promote or decrease protein solubility. Sci. Rep. *8*, 1–13.

Huang, K.-Y., Wu, H.-Y., Chen, Y.-J., Lu, C.-T., Su, M.-G., Hsieh, Y.-C., Tsai, C.-M., Lin, K.-I., Huang, H.-D., Lee, T.-Y., et al. (2014). RegPhos 2.0: an updated resource to explore protein kinase-substrate phosphorylation networks in mammals. Database J. Biol. Databases Curation *1*, 1–12.

Huang, S.Y., Shi, S.P., Qiu, J.D., and Liu, M.C. (2015). Using support vector machines to identify protein phosphorylation sites in viruses. J. Mol. Graph. Model. *56*, 84–90.

Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010). CD-HIT Suite: A web server for clustering and comparing biological sequences. Bioinformatics *26*, 680–682.

Iakoucheva, L.M., Radivojac, P., Brown, C.J., O'Connor, T.R., Sikes, J.G., Obradovic, Z., and Dunker, A.K. (2004). The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res. *32*, 1037–1049.

Ingrell, C.R., Miller, M.L., Jensen, O.N., and Blom, N. (2007). NetPhosYeast: Prediction of protein phosphorylation sites in yeast. Bioinformatics *23*, 895–897.

J.Kleywegt, G., and Jones, T.A. (1997). Model building and refinement practice. In Methods in Enzymology, pp. 208–231.

Jensen, O.N. (2004). Modification-specific proteomics: Characterization of post-translational modifications by mass spectrometry. Curr. Opin. Chem. Biol. *8*, 33–41.

Jiménez, J.L., Hegemann, B., Hutchins, J.R., Peters, J.M., and Durbin, R. (2007). A systematic comparative and structural analysis of protein phosphorylation sites based on the mtcPTM database. Genome Biol. *8*, 1–20.

Jung, I., Matsuyama, A., Yoshida, M., and Kim, D. (2010). PostMod: Sequence based prediction of kinase-specific phosphorylation sites with indirect relationship. BMC Bioinformatics *11*, 1–10.

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers *22*, 2577–2637.

Karabulut, N.P., and Frishman, D. (2016). Sequence- and structure-based analysis of tissue-specific phosphorylation sites. PLoS One *11*, 1–19.

Kessel, A., and Ben-Tal, N. (2018). Introduction to Proteins: Structure, Function, and Motion (CRC Press).

Klausen, M.S., Jespersen, M.C., Nielsen, H., Jensen, K.K., Jurtz, V.I., Soenderby, C.K., Sommer, M.O.A., Winther, O., Nielsen, M., Petersen, B., et al. NetSurfP-2.0: improved prediction of protein structural features by integrated deep learning. BioRxiv.

Kobe, B., Bod, M., Patrick, R., and Kim-anh, L. (2015). PhosphoPICK : modelling cellular context to map kinase-substrate phosphorylation events. Bioinformatics *31*, 382–389.

Koenig, M., and Grabe, N. (2004). Highly specific prediction of phosphorylation sites in proteins. Bioinformatics *20*, 3620–3627.

Konrat, R. (2009). The protein meta-structure: A novel concept for chemical and molecular biology. Cell. Mol. Life Sci. *66*, 3625–3639.

Krieger, E., and Vriend, G. (2014). YASARA View - molecular graphics for all devices - from

smartphones to workstations. Bioinformatics *30*, 2981–2982.

Landry, C.R., Levy, E.D., and Michnick, S.W. (2009). Weak functional constraints on phosphoproteomes. Trends Genet *25*, 193–197.

Lee, T.-Y., Huang, H.-D., Hung, J.-H., Huang, H.-Y., Yang, Y.-S., and Wang, T.-H. (2006). dbPTM: an information repository of protein post-translational modification. Nucleic Acids Res. *34*, 622–627.

Li, F., Li, C., Marquez-Lago, T.T., Leier, A., Akutsu, T., Purcell, A.W., Smith, A.I., Lithgow, T., Daly, R.J., Song, J., et al. (2018). Quokka: A comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome. Bioinformatics *34*, 4223–4231.

Li, H., Xing, X., Ding, G., Li, Q., Wang, C., Xie, L., Zeng, R., and Li, Y. (2009). SysPTM: a systematic resource for proteomic research on post-translational modifications. Mol. Cell. Proteomics *8*, 1839–1849.

Li, L., Wu, C., Huang, H., Zhang, K., Gan, J., and Li, S.S.C. (2008). Prediction of phosphotyrosine signaling networks using a scoring matrix-assisted ligand identification approach. Nucleic Acids Res. *36*, 3263–3273.

Mann, M., Kumar, C., Mijakovic, I., Olsen, J. V., Macek, B., Gnad, F., and Soufi, B. (2007). Phosphoproteome analysis of E. coli reveals evolutionary conservation of bacterial Ser/Thr/Tyr phosphorylation. Mol. Cell. Proteomics *7*, 299–307.

Manning, G., Whyte, D.B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002a). The protein kinase complement of the human genome. Science *298*, 1912–1934.

Manning, G., Plowman, G.D., Hunter, T., and Sudarsanam, S. (2002b). Evolution of protein kinase signaling from yeast to man. Trends Biochem. Sci. *27*, 514–520.

Miao, B., Xiao, Q., Chen, W., Li, Y., and Wang, Z. (2018). Evaluation of functionality for serine and threonine phosphorylation with different evolutionary ages in human and mouse. BMC Genomics *19*, 1–9.

Mishra, G.R., Suresh, M., Kumaran, K., Kannabiran, N., Suresh, S., Bala, P., Shivakumar, K., Anuradha, N., Reddy, R., Raghavan, T.M., et al. (2006). Human protein reference database--2006 update. Nucleic Acids Res. *34*, 411–414.

De Oliveira, P.S.L., Ferraz, F.A.N., Pena, D.A., Pramio, D.T., Morais, F.A., and Schechtman, D. (2016). Revisiting protein kinase-substrate interactions: Toward therapeutic development. Sci. Signal. *9*, 1–13.

Palmeri, A., Ferrè, F., and Helmer-Citterich, M. (2014). Exploiting holistic approaches to model specificity in protein phosphorylation. Front. Genet. *5*, 1–11.

Paul, M.K., and Mukhopadhyay, A.K. (2004). Tyrosine kinase - Role and significance in Cancer. Int. J. Med. Sci. *1*, 101–115.

Pearson, K. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. London, Edinburgh, Dublin Philos. Mag. J. Sci. *50*, 157–175.

Petersen, B., Petersen, T.N., Andersen, P., Nielsen, M., and Lundegaard, C. (2009). A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Struct. Biol. *9*, 1–10.

Plewczyński, D., Tkacz, A., Godzik, A., and Rychlewski, L. (2005). A support vector machine approach to the identification of phosphorylation sites .pdf. Cell. Mol. Biol. Lett. *10*, 73–89.

Radzicka, A., and Wolfenden, R. (1988). Comparing the polarities of the amino acids: Side-chain distribution coefficients between the vapor phase, cyclohexane, 1-Octanol, and neutral aqueous solution. Biochemistry *27*, 1664–1670.

Rajpurohit, Y.S., Bihani, S.C., Waldor, M.K., and Misra, H.S. (2016). Phosphorylation of Deinococcus radiodurans RecA regulates its activity and may contribute to radioresistance. J. Biol. Chem. *291*, 16672–16685.

Roach, P.J. (1991). Multisite and hierarchal protein phosphorylation. J. Biol. Chem. *266*, 14139–14142.

Sobolev, B., Filimonov, D., Lagunin, A., Zakharov, A., Koborova, O., Kel, A., and Poroikov, V. (2010). Functional classification of proteins based on projection of amino acid sequences: Application for prediction of protein kinase substrates. BMC Bioinformatics *11*, 1–18.

Stark, C., Su, T.-C., Breitkreutz, A., Lourenco, P., Dahabieh, M., Breitkreutz, B.-J., Tyers, M., and Sadowski, I. (2010). PhosphoGRID: a database of experimentally verified in vivo protein phosphorylation sites from the budding yeast Saccharomyces cerevisiae. Database (Oxford). *1*, 1–13.

Strumillo, M.J., Oplova, M., Vieitez, C., Ochoa, D., Shahraz, M., Busby, B.P., Sopko, R., Studer, R.A., Perrimon, N., Panse, V.G., et al. (2018). Conserved phosphorylation hotspots in eukaryotic protein domain families. BioRxiv *13*, 1–20.

Tan, C.S.H., Pasculescu, A., Lim, W.A., Pawson, T., Bader, G.D., and Linding, R. (2009). Positive selection of tyrosine loss in Metazoan evolution. Science *325*, 1686–1688.

Trost, B., and Kusalik, A. (2011). Computational prediction of eukaryotic phosphorylation sites. Bioinformatics *27*, 2927–2935.

Trost, B., and Kusalik, A. (2013). Computational phosphorylation site prediction in plants using random forests and organism-specific instance weights. Bioinformatics *29*, 686–694.

Trost, B., Kusalik, A., and Napper, S. (2016). Computational analysis of the predicted evolutionary conservation of human phosphorylation sites. PLoS One *11*, 1–14.

Tukey, J.W. (1949). Comparing individual means in the analysis of variance. Biometrics *5*, 99–114.

Tyanova, S., Cox, J., Olsen, J., Mann, M., and Frishman, D. (2013). Phosphorylation variation during the cell cycle scales with structural propensities of proteins. PLoS Comput. Biol. *9*, 1–10.

Ubersax, J.A., and Ferrell, J.E. (2007). Mechanisms of specificity in protein phosphorylation. Nat Rev Mol Cell Biol *8*, 530–541.

Velankar, S., Dana, J.M., Jacobsen, J., Van Ginkel, G., Gane, P.J., Luo, J., Oldfield, T.J., O'Donovan, C., Martin, M.J., and Kleywegt, G.J. (2013). SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Res. *41*, 483–489.

Vlastaridis, P., Kyriakidou, P., Chaliotis, A., Van de Peer, Y., Oliver, S.G., and Amoutzias, G.D. (2017). Estimating the total number of phosphoproteins and phosphorylation sites in eukaryotic proteomes. Gigascience *6*, 1–11.

Wang, G., and Dunbrack, R.L. (2005). PISCES: Recent improvements to a PDB sequence culling server. Nucleic Acids Res. *33*, 94–98.

Wang, D., Zeng, S., Xu, C., Qiu, W., Liang, Y., Joshi, T., and Xu, D. (2017). MusiteDeep: A deep-learning framework for general and kinase-specific phosphorylation site prediction. Bioinformatics *33*, 3909–3916.

Wang, J., Torii, M., Liu, H., Hart, G.W., and Hu, Z.-Z. (2011). dbOGAP - an integrated bioinformatics resource for protein O-GlcNAcylation. BMC Bioinformatics *12*, 91.

Wang, M.H., Li, C.H., Chen, W.Z., and Wang, C.X. (2008). Prediction of PK-specific phosphorylation site based on information entropy. Sci. China, Ser. C Life Sci. *51*, 12–20.

Wlodawer, A., Minor, W., Dauter, Z., Jaskolski, M., and Physics, B. (2008). Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. FEBS J. *275*, 1–21.

Wolfenden, R. V., Cullis, P.M., and Southgate, C.C.F. (1979). Water, protein folding, and the genetic code. Science (80-.). *206*, 575–577.

Xu, M., Fralick, D., Zheng, J.Z., Wang, B., Tu, X.M., and Feng, C. (2017). The differences and similarities between two-sample t-test and paired t-test. Shanghai Arch. Psychiatry *29*, 184–188.

Xue, Y., Li, A., Wang, L., Feng, H., and Yao, X. (2006). PPSP: Prediction of PK-specific phosphorylation site with Bayesian decision theory. BMC Bioinformatics *7*, 1–12.

Yang, C.-Y., Chang, C.-H., Yu, Y.-L., Emma Lin, T.-C., Lee, S.-A., Yen, C.-C., Yang, J.-M., Lai, J.-M., Hong, Y.-R., Tseng, T.-L., et al. (2008). PhosphoPOINT: A comprehensive human kinase interactome and phospho-protein database. Bioinformatics *24*, 14–20.

Yao, Q., Ge, H., Wu, S., Zhang, N., Chen, W., Xu, C., Gao, J., Thelen, J.J., and Xu, D. (2014). P^3DB 3.0: From plant phosphorylation sites to protein networks. Nucleic Acids Res. *42*, 1206–1213.

Yoo, P.D., Ho, Y.S., Zhou, B.B., and Zomaya, A.Y. (2008). SiteSeek: Post-translational modification analysis using adaptive locality-effective kernel methods and new profiles. BMC Bioinformatics *9*, 1–17.

Zanzoni, A., Carbajo, D., Diella, F., Gherardini, P.F., Tramontano, A., Helmer-Citterich, M., and Via, A. (2011). Phospho3D 2.0: An enhanced database of three-dimensional structures of phosphorylation sites. Nucleic Acids Res. *39*, 268–271.

Zilberstein, D., Tsigankov, P., Späth, G.F., Gherardini, P.F., Ausiello, G., Helmer-Citterich, M., and Palmeri, A. (2011). PhosTryp: a phosphorylation site predictor specific for parasitic protozoa of the family trypanosomatidae. BMC Genomics *12*, 614.

Supplementary material

Table 1: The number of the protein structures including pTyr less than 50 aa's long in categories based on protein folding ('with native-like fold', 'without secondary structure', 'including ladder as a part of beta sheet', or including alpha helix'). Protein chains including pTyr (cyan) are often crystalized as a part of protein complex (gray). Protein secondary structure of the chain including pTyr is marked out: dark blue alpha helices, green loops, and red ladders.

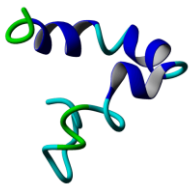
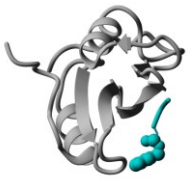
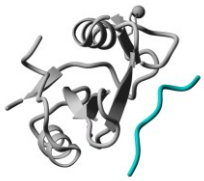
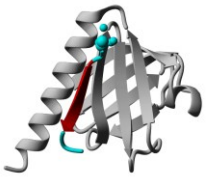
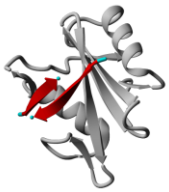
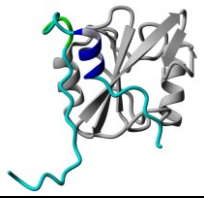
Short peptides:		**Image (PDB code)**	**Sums**
With native-like fold		2ljd 	7
Without secondary structure	**including other molecules**	1jyq 	47
	with unknown peptide backbone for phosphothreonine	3s3h 	2
	loop only	1kc2 	124
Including ladder as a part of beta sheet	**including other molecules**	5u1m 	3
	ladder only	1i3z 	29
	Forming beta sheet	3tkz 	1
Including alpha helix		2rsy 	1

Table 2: Distribution of (A) pP- and wP- sites (B) pP- and NonP- sites, and (C) wP- and NonP- sites in protein secondary structure elements (three categories according to DSSP (Kabsch and Sander, 1983): 'helix', 'strand', and 'loop'). For each amino acid dataset (serine = 'Ser', threonine = 'Thr', and tyrosine = 'Tyr') two subsets (30 and 90) were made, where the number corresponds to the maximum sequence identity shared by the protein chains in the redundancy set. Observed and chi-squared test expected frequencies are presented for each category. When the difference between distributions was significant at the 0.05 level, p-value was marked in bold.

(A)

dataset		helix		strand		loop		p-value
		pP	wP	pP	wP	pP	wP	
Tyr30	observed	7	10	4	2	20	11	0.2615
	expected	9.759	7.240	3.444	2.555	17.796	13.203	
Tyr90	observed	10	14	26	11	54	36	0.0836
	expected	14.304	9.695	22.052	14.947	53.642	36.357	
Ser30	observed	36	26	5	1	76	36	0.2851
	expected	40.178	21.821	3.888	2.111	71.932	39.067	
Ser90	observed	51	35	6	1	108	56	0.2862
	expected	55.214	30.785	4.494	2.505	105.291	58.708	
Thr30	observed	19	14	3	4	42	22	0.4269
	expected	20.307	12.692	4.307	2.692	39.384	24.615	
Thr90	observed	30	19	4	5	85	32	0.1059
	expected	33.320	15.680	6.120	2.880	79.560	37.440	

(B)

dataset		helix		strand		loop		p-value
		pP	NonP	pP	NonP	pP	NonP	
Tyr30	observed	7	177	4	90	20	115	**4.418×10^{-4}**
	expected	13.811	170.188	7.055	86.944	10.133	124.866	
Tyr90	observed	10	367	26	188	54	284	**5.366×10^{-9}**
	expected	36.523	340.476	20.731	193.268	32.744	305.255	
Ser30	observed	36	856	5	242	76	721	**2.410×10^{-7}**
	expected	53.907	838.092	14.927	232.072	48.165	748.834	
Ser90	observed	51	1 177	6	300	108	984	**3.964×10^{-10}**
	expected	77.159	1 150.840	19.226	286.773	68.613	1 023.386	
Thr30	observed	19	349	3	165	42	379	**5.167×10^{-4}**
	expected	24.610	343.389	11.235	156.764	28.154	392.845	
Thr90	observed	30	635	4	257	85	700	**1.384×10^{-8}**
	expected	46.250	618.749	18.152	242.847	54.596	730.403	

(C)

dataset		helix		strand		loop		p-value
		wP	NonP	wP	NonP	wP	NonP	
Tyr30	observed	10	177	2	90	11	115	0.1143
	expected	10.619	176.380	5.224	86.775	7.155	118.844	
Tyr90	observed	14	367	11	188	36	284	**2.707×10^{-4}**
	expected	25.823	355.176	13.487	185.512	21.688	298.311	
Ser30	observed	26	856	1	242	36	721	**3.106×10^{-3}**
	expected	29.524	852.475	8.134	234.865	25.340	731.659	
Ser90	observed	35	1 177	1	300	56	984	**3.435×10^{-5}**
	expected	43.675	1 168.324	10.846	290.153	37.477	1 002.522	
Thr30	observed	14	349	4	165	22	379	0.2136
	expected	15.562	347.437	7.245	161.754	17.191	383.808	
Thr90	observed	19	635	5	257	32	700	0.1126
	expected	22.223	631.7767	8.902	253.097	24.873	707.126	

Table 3: The average distances between the sites and aa's neighboring the sites selected by Euclidean distance and Voronoi diagrams. For each amino acid dataset (serine = 'Ser', threonine = 'Thr', and tyrosine = 'Tyr') two subsets (30 and 90) were made, where the number corresponds to the maximum sequence identity shared by the protein chains in the redundancy set.

dataset		Euclidean	Voronoi
Tyr30	pP	8.228	5.540
	wP	8.163	5.914
	NonP	8.408	5.795
Tyr90	pP	7.701	5.920
	wP	8.124	17.531
	NonP	8.322	5.830
Ser30	pP	6.855	7.567
	wP	5.999	7.864
	NonP	6.136	6.033
Ser90	pP	6.675	7.049
	wP	6.044	7.142
	NonP	6.107	5.904
Thr30	pP	6.604	6.061
	wP	6.176	5.944
	NonP	6.218	5.712
Thr90	pP	6.730	6.275
	wP	6.212	5.947
	NonP	6.202	5.685

Table 4: Results of multiple ANOVA analysis of the sites compactness. The means of average distance between sites and aa's neighboring the sites selected by Euclidean distance and Voronoi diagrams were calculated and compared between datasets (datasets1-datasets2). Mean difference between datasets, standard error ('Std. Error'), significance ('Sig.'), and 95% confidence interval are presented. ANOVA analysis was made for (A) tyrosine sites, (B) serine sites, and (C) threonine sites. For each amino acid dataset (serine = 'Ser', threonine = 'Thr', and tyrosine = 'Tyr') two subsets (30 and 90) were made, where the number corresponds to the maximum sequence identity shared by the protein chains in the redundancy set.

(A)

Multiple Comparisons (Tukey HSD)																				
Method	dataset1	dataset2	Mean Difference (1-2)	Std. Error	Sig.	95% Confidence Interval		dataset1	dataset2	Mean Difference (1-2)	Std. Error	Sig.	95% Confidence Interval		dataset1	dataset2	Mean Difference (1-2)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound						Lower Bound	Upper Bound						Lower Bound	Upper Bound
Euclidean	Tyr30pP	Ser30NonP	2.092*	0.221	7.525×10^{-12}	1.320	2.865	Tyr30wP	Ser30NonP	2.027*	0.262	9.263×10^{-12}	1.112	2.942	Tyr30NonP	Ser30NonP	2.272*	0.066	7.525×10^{-12}	2.041	2.503
		Ser30pP	1.373*	0.247	4.387×10^{-6}	0.509	2.236		Ser30pP	1.307*	0.284	6.079×10^{-4}	0.315	2.300		Ser30pP	1.552*	0.128	7.525×10^{-12}	1.104	2.001
		Ser30wP	2.229*	0.268	7.541×10^{-12}	1.293	3.164		Ser30wP	2.163*	0.302	1.508×10^{-10}	1.107	3.220		Ser30wP	2.408*	0.165	7.525×10^{-12}	1.832	2.985
		Thr30NonP	2.009*	0.223	7.525×10^{-12}	1.230	2.789		Thr30NonP	1.944*	0.263	3.516×10^{-11}	1.024	2.865		Thr30NonP	2.189*	0.072	7.525×10^{-12}	1.937	2.441
		Thr30pP	1.624*	0.268	2.198×10^{-7}	0.688	2.559		Thr30pP	1.558*	0.302	3.934×10^{-5}	0.502	2.615		Thr30pP	1.803*	0.165	7.525×10^{-12}	1.227	2.379
		Thr30wP	2.052*	0.292	3.740×10^{-10}	1.032	3.073		Thr30wP	1.987*	0.324	1.424×10^{-7}	0.855	3.119		Thr30wP	2.232*	0.202	7.525×10^{-12}	1.526	2.938
		Tyr30NonP	-0.179	0.227	0.999	-0.973	0.613		Tyr30NonP	-0.244	0.267	0.999	-1.177	0.687		Tyr30pP	0.179	0.227	0.999	-0.613	0.973
		Tyr30wP	0.065	0.340	1.000	-1.123	1.254		Tyr30pP	-0.065	0.340	1.000	-1.254	1.123		Tyr30wP	0.244	0.267	0.999	-0.687	1.177
	Tyr90pP	Ser90NonP	1.593*	0.131	7.525×10^{-12}	1.133	2.054	Tyr90wP	Ser90NonP	2.017*	0.162	7.525×10^{-12}	1.450	2.583	Tyr90NonP	Ser90NonP	2.215*	0.048	7.525×10^{-12}	2.047	2.383
		Ser90pP	1.025*	0.160	2.948×10^{-08}	0.464	1.587		Ser90pP	1.448*	0.186	8.954×10^{-12}	0.797	2.100		Ser90pP	1.647*	0.103	7.525×10^{-12}	1.284	2.010
		Ser90wP	1.656*	0.181	7.525×10^{-12}	1.022	2.290		Ser90wP	2.079*	0.204	7.525×10^{-12}	1.364	2.794		Ser90wP	2.277*	0.133	7.525×10^{-12}	1.810	2.745
		Thr90NonP	1.498*	0.133	7.525×10^{-12}	1.034	1.963		Thr90NonP	1.922*	0.163	7.525×10^{-12}	1.351	2.492		Thr90NonP	2.120*	0.051	7.525×10^{-12}	1.941	2.299
		Thr90pP	0.970*	0.171	2.434×10^{-06}	0.371	1.569		Thr90pP	1.393*	0.195	1.982×10^{-10}	0.709	2.077		Thr90pP	1.591*	0.119	7.525×10^{-12}	1.173	2.010
		Thr90wP	1.488*	0.207	1.238×10^{-10}	0.764	2.211		Thr90wP	1.911*	0.227	7.534×10^{-12}	1.116	2.707		Thr90wP	2.110*	0.167	7.525×10^{-12}	1.527	2.693
		Tyr90NonP	-0.621*	0.135	6.892×10^{-4}	-1.096	-0.147		Tyr90NonP	-0.198	0.165	0.999	-0.776	0.379		Tyr90pP	0.621*	0.135	6.892×10^{-4}	0.147	1.096
		Tyr90wP	-0.423	0.206	0.843	-1.143	0.296		Tyr90pP	0.423	0.206	0.843	-0.296	1.143		Tyr90wP	0.198	0.165	0.999	-0.379	0.776
Voronoi	Tyr30pP	Ser30NonP	-0.492	0.877	1.000	-3.556	2.571	Tyr30wP	Ser30NonP	-0.118	1.033	1.000	-3.723	3.486	Tyr30NonP	Ser30NonP	-0.238	0.265	1.000	-1.165	0.689
		Ser30pP	-2.026	0.982	0.837	-5.453	1.401		Ser30pP	-1.652	1.122	0.993	-5.571	2.266		Ser30pP	-1.771	0.514	0.058	-3.566	0.022
		Ser30wP	-2.324	1.068	0.771	-6.053	1.405		Ser30wP	-1.949	1.199	0.979	-6.135	2.235		Ser30wP	-2.069	0.664	0.151	-4.389	0.250
		Thr30NonP	-0.171	0.885	1.000	-3.262	2.919		Thr30NonP	0.202	1.039	1.000	-3.425	3.831		Thr30NonP	0.083	0.290	1.000	-0.930	1.097
		Thr30pP	-0.520	1.065	1.000	-4.240	3.199		Thr30pP	-0.146	1.196	1.000	-4.323	4.030		Thr30pP	-0.266	0.660	1.000	-2.570	2.038
		Thr30wP	-0.403	1.167	1.000	-4.478	3.671		Thr30wP	-0.029	1.288	1.000	-4.525	4.466		Thr30wP	-0.148	0.814	1.000	-2.990	2.693
		Tyr30NonP	-0.254	0.902	1.000	-3.404	2.895		Tyr30NonP	0.119	1.054	1.000	-3.559	3.798		Tyr30pP	0.254	0.902	1.000	-2.895	3.404
		Tyr30wP	-0.374	1.345	1.000	-5.070	4.322		Tyr30pP	0.374	1.345	1.000	-4.322	5.070		Tyr30wP	-0.119	1.054	1.000	-3.798	3.559
	Tyr90pP	Ser90NonP	0.015	0.525	1.000	-1.818	1.849	Tyr90wP	Ser90NonP	11.626*	0.638	0.000	9.399	13.852	Tyr90NonP	Ser90NonP	-0.074	0.193	1.000	-0.748	0.599
		Ser90pP	-1.129	0.642	0.955	-3.370	1.111		Ser90pP	10.481*	0.737	0.000	7.908	13.053		Ser90pP	-1.219	0.416	0.238	-2.673	0.234
		Ser90wP	-1.222	0.727	0.971	-3.762	1.317		Ser90wP	10.388*	0.812	0.000	7.551	13.225		Ser90wP	-1.312	0.539	0.583	-3.194	0.570
		Thr90NonP	0.234	0.530	1.000	-1.617	2.086		Thr90NonP	11.845*	0.642	0.000	9.603	14.087		Thr90NonP	0.144	0.206	1.000	-0.577	0.866
		Thr90pP	-0.354	0.685	1.000	-2.747	2.037		Thr90pP	11.256*	0.775	0.000	8.550	13.961		Thr90pP	-0.444	0.480	1.000	-2.122	1.233
		Thr90wP	-0.027	0.831	1.000	-2.930	2.874		Thr90wP	11.583*	0.907	0.000	8.417	14.748		Thr90wP	-0.117	0.672	1.000	-2.466	2.230
		Tyr90NonP	0.089	0.542	1.000	-1.802	1.981		Tyr90NonP	11.700*	0.651	0.000	9.426	13.975		Tyr90pP	-0.089	0.542	1.000	-1.981	1.802
		Tyr90wP	-11.610*	0.814	0.000	-14.454	-8.767		Tyr90pP	11.610*	0.814	0.000	8.767	14.454		Tyr90wP	-11.700*	0.651	0.000	-13.975	-9.426

(B)

	Multiple Comparisons (Tukey HSD)																				
Method	dataset1	dataset2	Mean Difference (1-2)	Std. Error	Sig.	95% Confidence Interval		dataset1	dataset2	Mean Difference (1-2)	Std. Error	Sig.	95% Confidence Interval		dataset1	dataset2	Mean Difference (1-2)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound						Lower Bound	Upper Bound						Lower Bound	Upper Bound
Euclidean	Ser30pP	Ser30NonP	0.719*	0.117	1.443×10^{-7}	0.309	1.129	Ser30wP	Ser30NonP	-0.136	0.156	0.999	-0.682	0.410	Ser30NonP	Ser30pP	-0.719*	0.117	1.443×10^{-7}	-1.129	-0.309
		Ser30wP	0.856*	0.191	1.105×10^{-3}	0.187	1.524		Ser30pP	-0.856*	0.191	0.001	-1.524	-0.187		Ser30wP	0.136	0.156	0.999	-0.410	0.682
		Thr30NonP	0.636*	0.121	2.202×10^{-5}	0.214	1.059		Thr30NonP	-0.219	0.159	0.996	-0.775	0.336		Thr30NonP	-0.082	0.049	0.973	-0.257	0.091
		Thr30pP	0.250	0.191	0.998	-0.417	0.919		Thr30pP	-0.605	0.217	0.326	-1.365	0.154		Thr30pP	-0.468	0.156	0.203	-1.015	0.077
		Thr30wP	0.679	0.224	0.186	-0.103	1.462		Thr30wP	-0.176	0.247	0.999	-1.038	0.685		Thr30wP	-0.040	0.195	1.000	-0.721	0.641
		Tyr30NonP	-1.552*	0.128	7.525×10^{-12}	-2.001	-1.104		Tyr30NonP	-2.408*	0.165	7.525×10^{-12}	-2.985	-1.832		Tyr30NonP	-2.272*	0.066	7.525×10^{-12}	-2.503	-2.041
		Tyr30pP	-1.373*	0.247	4.387×10^{-6}	-2.236	-0.509		Tyr30pP	-2.229*	0.268	7.541×10^{-12}	-3.164	-1.293		Tyr30pP	-2.092*	0.221	7.525×10^{-12}	-2.865	-1.320
		Tyr30wP	-1.307*	0.284	6.079×10^{-4}	-2.300	-0.315		Tyr30wP	-2.163*	0.302	1.508×10^{-10}	-3.220	-1.107		Tyr30wP	-2.027*	0.262	9.263×10^{-12}	-2.942	-1.112
	Ser90pP	Ser90NonP	0.568*	0.098	1.263×10^{-6}	0.224	0.912	Ser90wP	Ser90NonP	-0.062	0.129	0.999	-0.515	0.390	Ser90NonP	Ser90pP	-0.568*	0.098	1.263×10^{-6}	-0.912	-0.224
		Ser90wP	0.630*	0.159	9.334×10^{-3}	0.075	1.186		Ser90pP	-0.630*	0.159	0.009	-1.186	-0.075		Ser90wP	0.062	0.129	0.999	-0.390	0.515
		Thr90NonP	0.473*	0.100	3.411×10^{-4}	0.123	0.822		Thr90NonP	-0.157	0.131	0.999	-0.614	0.299		Thr90NonP	-0.095	0.039	0.590	-0.231	0.041
		Thr90pP	-0.055	0.147	0.999	-0.570	0.459		Thr90pP	-0.685*	0.169	6.962×10^{-3}	-1.279	-0.092		Thr90pP	-0.623*	0.115	9.391×10^{-6}	-1.025	-0.221
		Thr90wP	0.462	0.187	0.559	-0.193	1.118		Thr90wP	-0.167	0.206	0.999	-0.886	0.551		Thr90wP	-0.105	0.163	0.999	-0.676	0.465
		Tyr90NonP	-1.647*	0.103	7.525×10^{-12}	-2.010	-1.284		Tyr90NonP	-2.277*	0.133	7.525×10^{-12}	-2.745	-1.810		Tyr90NonP	-2.215*	0.048	7.525×10^{-12}	-2.383	-2.047
		Tyr90pP	-1.025*	0.160	2.948×10^{-8}	-1.587	-0.464		Tyr90pP	-1.656*	0.181	7.525×10^{-12}	-2.290	-1.022		Tyr90pP	-1.593*	0.131	7.525×10^{-12}	-2.054	-1.133
		Tyr90wP	-1.448*	0.186	8.954×10^{-12}	-2.100	-0.797		Tyr90wP	-2.079*	0.204	7.525×10^{-12}	-2.794	-1.364		Tyr90wP	-2.017*	0.162	7.525×10^{-12}	-2.583	-1.450
Voronoi	Ser30pP	Ser30NonP	1.533	0.469	0.099	-0.105	3.172	Ser30wP	Ser30NonP	1.831	0.630	0.251	-0.370	4.033	Ser30NonP	Ser30pP	-1.533	0.469	0.099	-3.172	0.105
		Ser30wP	-0.297	0.769	1.000	-2.98	2.387		Ser30pP	0.297	0.769	1.000	-2.387	2.982		Ser30wP	-1.831	0.630	0.251	-4.033	0.370
		Thr30NonP	1.854*	0.484	0.015	0.165	3.544		Thr30NonP	2.152	0.641	0.076	-0.086	4.392		Thr30NonP	0.321	0.201	0.982	-0.380	1.023
		Thr30pP	1.505	0.765	0.885	-1.165	4.176		Thr30pP	1.803	0.873	0.837	-1.245	4.852		Thr30pP	-0.027	0.626	1.000	-2.213	2.157
		Thr30wP	1.622	0.901	0.944	-1.524	4.769		Thr30wP	1.920	0.995	0.901	-1.552	5.394		Thr30wP	0.089	0.786	1.000	-2.656	2.835
		Tyr30NonP	1.771	0.514	0.058	-0.022	3.566		Tyr30NonP	2.069	0.664	0.151	-0.250	4.389		Tyr30NonP	0.238	0.265	1.000	-0.689	1.165
		Tyr30pP	2.026	0.982	0.837	-1.401	5.453		Tyr30pP	2.324	1.068	0.771	-1.405	6.053		Tyr30pP	0.492	0.877	1.000	-2.571	3.556
		Tyr30wP	1.652	1.122	0.993	-2.266	5.571		Tyr30wP	1.949	1.199	0.979	-2.235	6.135		Tyr30wP	0.118	1.033	1.000	-3.486	3.723
	Ser90pP	Ser90NonP	1.144	0.394	0.252	-0.232	2.522	Ser90wP	Ser90NonP	1.237	0.522	0.633	-0.586	3.061	Ser90NonP	Ser90pP	-1.144	0.394	0.252	-2.522	0.232
		Ser90wP	-0.092	0.639	1.000	-2.325	2.140		Ser90pP	0.092	0.639	1.000	-2.140	2.325		Ser90wP	-1.235	0.522	0.633	-3.061	0.586
		Thr90NonP	1.364	0.401	0.067	-0.037	2.765		Thr90NonP	1.456	0.527	0.339	-0.385	3.299		Thr90NonP	0.219	0.158	0.996	-0.332	0.770
		Thr90pP	0.774	0.591	0.998	-1.289	2.838		Thr90pP	0.867	0.683	0.999	-1.517	3.252		Thr90pP	-0.370	0.462	1.000	-1.982	1.242
		Thr90wP	1.101	0.755	0.993	-1.535	3.739		Thr90wP	1.194	0.829	0.994	-1.701	4.090		Thr90wP	-0.043	0.659	1.000	-2.344	2.258
		Tyr90NonP	1.219	0.416	0.238	-0.234	2.673		Tyr90NonP	1.312	0.539	0.583	-0.570	3.194		Tyr90NonP	0.074	0.193	1.000	-0.599	0.748
		Tyr90pP	1.129	0.642	0.955	-1.111	3.370		Tyr90pP	1.222	0.727	0.971	-1.317	3.762		Tyr90pP	-0.015	0.525	1.000	-1.849	1.818
		Tyr90wP	-10.481*	0.737	0.000	-13.053	-7.908		Tyr90wP	-10.388*	0.812	0.000	-13.225	-7.551		Tyr90wP	-11.626*	0.638	0.000	-13.852	-9.399

(c)

Multiple Comparisons (Tukey HSD)																				
Method	dataset1	dataset2	Mean Difference (1-2)	Std. Error	Sig.	95% Confidence Interval		dataset1	dataset2	Mean Difference (1-2)	Std. Error	Sig.	95% Confidence Interval		dataset1	dataset2	Mean Difference (1-2)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound						Lower Bound	Upper Bound						Lower Bound	Upper Bound
Euclidean	Thr30pP	Ser30NonP	0.468	0.156	0.203	-0.077	1.015	Thr30wP	Ser30NonP	0.040	0.195	1.000	-0.641	0.721	Thr30NonP	Ser30NonP	0.082	0.049	0.973	-0.091	0.257
		Ser30pP	-0.250	0.191	0.998	-0.919	0.417		Ser30pP	-0.679	0.224	0.186	-1.462	0.103		Ser30pP	-0.6368*	0.121	2.202×10^{-5}	-1.059	-0.214
		Ser30wP	0.605	0.217	0.326	-0.154	1.365		Ser30wP	0.176	0.247	0.999	-0.685	1.038		Ser30wP	0.219	0.159	0.996	-0.336	0.775
		Thr30NonP	0.385	0.159	0.590	-0.170	0.941		Thr30NonP	-0.042	0.197	0.999	-0.732	0.646		Thr30pP	-0.385	0.159	0.590	-0.941	0.170
		Thr30wP	0.428	0.247	0.960	-0.433	1.290		Thr30pP	-0.428	0.247	0.960	-1.290	0.433		Thr30wP	0.042	0.197	0.999	-0.646	0.732
		Tyr30NonP	-1.803*	0.165	7.525×10^{-12}	-2.379	-1.227		Tyr30NonP	-2.232*	0.202	7.525×10^{-12}	-2.938	-1.526		Tyr30NonP	-2.189*	0.072	7.525×10^{-12}	-2.441	-1.937
		Tyr30pP	-1.624*	0.268	2.198×10^{-7}	-2.559	-0.688		Tyr30pP	-2.052*	0.292	3.740×10^{-10}	-3.073	-1.032		Tyr30pP	-2.009*	0.223	7.525×10^{-12}	-2.789	-1.230
		Tyr30wP	-1.558*	0.302	3.934×10^{-5}	-2.615	-0.502		Tyr30wP	-1.987*	0.324	1.424×10^{-7}	-3.119	-0.855		Tyr30wP	-1.944*	0.263	3.516×10^{-11}	-2.865	-1.024
	Thr90pP	Ser90NonP	0.623*	0.115	9.391×10^{-6}	0.221	1.025	Thr90wP	Ser90NonP	0.105	0.163	0.999	-0.465	0.676	Thr90NonP	Ser90NonP	0.095	0.039	0.590	-0.041	0.231
		Ser90pP	0.055	0.147	0.999	-0.459	0.570		Ser90pP	-0.462	0.187	0.559	-1.118	0.193		Ser90pP	-0.473*	0.100	3.411×10^{-4}	-0.822	-0.123
		Ser90wP	0.685*	0.169	6.962×10^{-3}	0.092	1.279		Ser90wP	0.167	0.206	0.999	-0.551	0.886		Ser90wP	0.157	0.131	0.999	-0.299	0.614
		Thr90NonP	0.528*	0.116	8.239×10^{-4}	0.121	0.935		Thr90NonP	0.010	0.164	1.000	-0.564	0.585		Thr90pP	-0.528*	0.116	8.239×10^{-4}	-0.935	-0.121
		Thr90wP	0.518	0.197	0.432	-0.169	1.206		Thr90pP	-0.518	0.197	0.432	-1.206	0.169		Thr90wP	-0.010	0.164	1.000	-0.585	0.564
		Tyr90NonP	-1.591*	0.119	7.525×10^{-12}	-2.010	-1.173		Tyr90NonP	-2.110*	0.167	7.525×10^{-12}	-2.693	-1.527		Tyr90NonP	-2.120*	0.051	7.525×10^{-12}	-2.299	-1.941
		Tyr90pP	-0.970*	0.171	2.434×10^{-6}	-1.569	-0.371		Tyr90pP	-1.488*	0.207	1.238×10^{-10}	-2.211	-0.764		Tyr90pP	-1.498*	0.133	7.525×10^{-12}	-1.963	-1.034
		Tyr90wP	-1.393*	0.195	1.982×10^{-10}	-2.077	-0.709		Tyr90wP	-1.911*	0.227	7.534×10^{-12}	-2.707	-1.116		Tyr90wP	-1.922*	0.163	7.525×10^{-12}	-2.492	-1.351
Voronoi	Thr30pP	Ser30NonP	0.027	0.626	1.000	-2.157	2.213	Thr30wP	Ser30NonP	-0.089	0.786	1.000	-2.835	2.656	Thr30NonP	Ser30NonP	-0.321	0.201	0.982	-1.023	0.380
		Ser30pP	-1.505	0.765	0.885	-4.176	1.165		Ser30pP	-1.622	0.901	0.944	-4.769	1.524		Ser30pP	-1.854*	0.484	0.015	-3.544	-0.165
		Ser30wP	-1.803	0.873	0.837	-4.852	1.245		Ser30wP	-1.920	0.995	0.901	-5.394	1.552		Ser30wP	-2.152	0.641	0.076	-4.392	0.086
		Thr30NonP	0.349	0.637	1.000	-1.873	2.572		Thr30NonP	0.232	0.795	1.000	-2.544	3.008		Thr30pP	-0.349	0.637	1.000	-2.572	1.873
		Thr30wP	0.117	0.992	1.000	-3.345	3.580		Thr30pP	-0.117	0.992	1.000	-3.580	3.345		Thr30wP	-0.232	0.795	1.000	-3.008	2.544
		Tyr30NonP	0.266	0.660	1.000	-2.038	2.570		Tyr30NonP	0.148	0.814	1.000	-2.693	2.990		Tyr30NonP	-0.083	0.290	1.000	-1.097	0.930
		Tyr30pP	0.520	1.065	1.000	-3.199	4.240		Tyr30pP	0.403	1.167	1.000	-3.671	4.478		Tyr30pP	0.171	0.885	1.000	-2.919	3.262
		Tyr30wP	0.146	1.196	1.000	-4.030	4.323		Tyr30wP	0.029	1.288	1.000	-4.466	4.525		Tyr30wP	-0.202	1.039	1.000	-3.831	3.425
	Thr90pP	Ser90NonP	0.370	0.462	1.000	-1.242	1.982	Thr90wP	Ser90NonP	0.043	0.659	1.000	-2.258	2.344	Thr90NonP	Ser90NonP	-0.219	0.158	0.996	-0.770	0.332
		Ser90pP	-0.774	0.591	0.998	-2.838	1.289		Ser90pP	-1.101	0.755	0.993	-3.739	1.535		Ser90pP	-1.364	0.401	0.067	-2.765	0.037
		Ser90wP	-0.867	0.683	0.999	-3.252	1.517		Ser90wP	-1.194	0.829	0.994	-4.090	1.701		Ser90wP	-1.456	0.527	0.339	-3.299	0.385
		Thr90NonP	0.589	0.467	0.999	-1.043	2.222		Thr90NonP	0.262	0.663	1.000	-2.053	2.578		Thr90pP	-0.589	0.467	0.999	-2.222	1.043
		Thr90wP	0.327	0.793	1.000	-2.440	3.094		Thr90pP	-0.327	0.793	1.000	-3.094	2.440		Thr90wP	-0.262	0.663	1.000	-2.575	2.053
		Tyr90NonP	0.444	0.480	1.000	-1.233	2.122		Tyr90NonP	0.117	0.672	1.000	-2.230	2.466		Tyr90NonP	-0.144	0.206	1.000	-0.866	0.577
		Tyr90pP	0.354	0.685	1.000	-2.037	2.747		Tyr90pP	0.027	0.831	1.000	-2.874	2.930		Tyr90pP	-0.234	0.530	1.000	-2.086	1.617
		Tyr90wP	-11.256*	0.775	0.000	-13.965	-8.550		Tyr90wP	-11.583*	0.907	0.000	-14.748	-8.417		Tyr90wP	-11.845*	0.642	0.000	-14.087	-9.603

Table 5: The distribution of (A) pP- and wP-sites, (B) pP- and NonP-sites, and (C) wP- and NonP-sites located on the protein surface ('exposed') and within protein structure ('buried'). For each amino acid dataset (serine = 'Ser', threonine = 'Thr', and tyrosine = 'Tyr') two subsets (30 and 90) were made, where the number corresponds to the maximum sequence identity shared by the protein chains in the redundancy set. Observed and chi-squared test expected frequencies are presented. When the difference between distributions was significant at the 0.05 level, p-value was marked in bold.

(A)

dataset		exposed		buried		p-value
		pP	wP	pP	wP	
Tyr30	observed	16	10	16	13	0.6328
	expected	15.127	10.872	16.872	12.127	
Tyr90	observed	59	37	32	24	0.6006
	expected	57.473	38.526	33.526	22.473	
Ser30	observed	87	41	30	22	0.1901
	expected	83.200	44.800	33.800	18.200	
Ser90	observed	125	54	41	38	**5.574×10^{-3}**
	expected	115.170	63.829	50.829	28.170	
Thr30	observed	48	27	16	13	0.4067
	expected	46.153	28.846	17.846	11.153	
Thr90	observed	75	34	44	23	0.6660
	expected	73.698	35.301	45.301	21.698	

(B)

dataset		exposed		buried		p-value
		pP	NonP	pP	NonP	
Tyr30	observed	16	107	16	316	**2.414×10^{-3}**
	expected	8.650	114.349	23.349	308.650	
Tyr90	observed	59	248	32	316	**8.147×10^{-13}**
	expected	28.771	278.228	62.228	601.771	
Ser30	observed	87	1 106	30	713	**3.470×10^{-3}**
	expected	72.097	1 120.902	44.902	698.097	
Ser90	observed	125	1 507	41	978	**1.708×10^{-4}**
	expected	102.192	1 529.807	63.807	955.192	
Thr30	observed	48	478	16	415	**8.523×10^{-4}**
	expected	35.176	490.823	28.823	402.176	
Thr90	observed	75	869	44	723	7.414×10^{-2}
	expected	65.655	878.344	53.344	713.655	

(c)

dataset		exposed		buried		p-value
		wP	NonP	wP	NonP	
Tyr30	observed	10	107	13	316	5.354×10^{-2}
	expected	6.033	110.966	16.966	312.033	
Tyr90	observed	37	248	24	632	**9.409×10^{-8}**
	expected	18.475	266.524	42.524	613.475	
Ser30	observed	41	1 106	22	713	0.4939
	expected	38.395	1 108.604	24.604	710.395	
Ser90	observed	54	1 507	38	978	0.7073
	expected	55.728	1 505.271	36.271	979.728	
Thr30	observed	27	478	13	415	8.274×10^{-2}
	expected	21.650	483.349	18.349	409.650	
Thr90	observed	34	869	23	723	0.4504
	expected	31.213	871.786	25.786	720.213	

Table 6: Distribution of aa's spatially neighboring the pP-, wP-, and NonP-sites in three groups based on hydrophobicity (hydrophobic = 'phobic', hydrophilic = 'philic', and 'polar)'. Residues that were not from the list of 20 'canonical 'amino acids, were treated as 'hetatm' (red) and excluded from further analysis. The distribution is presented as the number ('count') of aa's spatially neighboring the site and a relative percentage ('percent'), where sum of all aa'a in particular dataset is 100%. For each amino acid dataset (serine = 'Ser', threonine = 'Thr', and tyrosine = 'Tyr') two subsets (30 and 90) were made, where the number corresponds to the maximum sequence identity shared by the protein chains in the redundancy set. Amino acids spatially neighboring the sites were selected by Euclidean distance and Voronoi diagrams.

dataset	area	type	count					percent		
			phobic	philic	polar	hetatm	sum aa's	phobic	philic	polar
Tyr30	Euclidean	pP	76	108	38	55	222	34.234	48.648	17.117
		wP	79	54	33	34	166	47.590	32.530	19.879
		NonP	1 701	1 207	504	886	3 412	49.853	35.375	14.771
Tyr30	Voronoi	pP	106	65	39	34	210	50.476	30.952	18.571
		wP	87	63	37	23	187	46.524	33.689	19.786
		NonP	1 848	1 147	538	401	3 533	52.306	32.465	15.227
Tyr90	Euclidean	pP	134	260	96	152	490	27.346	53.061	19.591
		wP	182	171	67	131	420	43.333	40.714	15.952
		NonP	3 571	2 454	832	1 588	6 857	52.078	35.788	12.133
Tyr90	Voronoi	pP	254	207	165	107	626	40.575	33.067	26.357
		wP	201	173	77	76	451	44.567	38.359	17.073
		NonP	3 900	2 381	992	789	7 273	53.622	32.737	13.639
Ser30	Euclidean	pP	378	443	146	276	967	39.089	45.811	15.098
		wP	204	216	94	141	514	39.688	42.023	18.287
		NonP	8 742	5 918	2 531	3 687	17 191	50.852	34.424	14.722
Ser30	Voronoi	pP	294	286	103	119	683	43.045	41.874	15.080
		wP	189	155	63	60	407	46.437	38.083	15.479
		NonP	6 041	4 165	1 787	1 726	11 993	50.371	34.728	14.900
Ser90	Euclidean	pP	527	604	194	366	1 325	39.773	45.584	14.641
		wP	301	320	132	225	753	39.973	42.496	17.529
		NonP	11 864	8 114	3 363	5 204	23 341	50.829	34.762	14.408
Ser90	Voronoi	pP	434	387	141	157	962	45.114	40.228	14.656
		wP	278	216	93	93	587	47.359	36.797	15.843
		NonP	8 243	5 722	2 363	2 384	16 328	50.483	35.044	14.472
Thr30	Euclidean	pP	194	252	73	126	519	37.379	48.554	14.065
		wP	129	123	105	56	357	36.134	34.453	29.411
		NonP	4 544	2 975	1 298	1 558	8 817	51.536	33.741	14.721
Thr30	Voronoi	pP	164	142	50	51	356	46.067	39.887	14.044
		wP	116	91	41	22	248	46.774	36.693	16.532
		NonP	3 072	2 051	810	663	5 933	51.778	34.569	13.652
Thr90	Euclidean	pP	331	451	146	245	928	35.668	48.599	15.732
		wP	171	183	92	110	446	38.340	41.031	20.627
		NonP	7 668	5 367	2 212	2 860	15 247	50.291	35.200	14.507
Thr90	Voronoi	pP	306	261	108	113	675	45.333	38.666	16.000
		wP	156	133	62	39	351	44.444	37.891	17.663
		NonP	5 271	3 696	1 424	1 251	10 391	50.726	35.569	13.704

Table 7: The distribution of aa's spatially neighboring (A) pP- and wP-sites, (B) pP- and NonP-sites, and (C) wP- and NonP-sites in three groups based on hydrophobicity (hydrophobic = 'phobic', hydrophilic = 'philic', and 'polar)'. For each amino acid dataset (serine = 'Ser', threonine = 'Thr', and tyrosine = 'Tyr') two subsets (30 and 90) were made, where the number corresponds to the maximum sequence identity shared by the protein chains in the redundancy set. Observed and chi-squared test expected frequencies are presented. When the difference between distributions was significant at the 0.05 level, p-value was marked in bold.

(A)

dataset	area	type	phobic		philic		polar		p-value
			pP	wP	pP	wP	pP	wP	
Tyr30	**Euclidean**	**observed**	76	79	108	54	38	33	**5.125×10^{-3}**
		expected	88.685	66.314	92.690	69.309	40.623	30.376	
Tyr30	**Voronoi**	**observed**	106	87	65	63	39	37	0.7320
		expected	102.090	90.909	67.707	60.292	40.201	35.798	
Tyr90	**Euclidean**	**observed**	134	182	260	171	96	67	**2.767×10^{-6}**
		expected	170.153	145.846	232.076	198.923	87.769	75.230	
Tyr90	**Voronoi**	**observed**	254	201	207	173	165	77	**1.411×10^{-3}**
		expected	264.466	190.533	220.872	159.127	140.661	101.338	
Ser30	**Euclidean**	**observed**	378	204	443	216	146	94	0.1999
		expected	380.009	201.990	430.285	228.714	156.704	83.295	
Ser30	**Voronoi**	**observed**	294	189	286	155	103	63	0.4506
		expected	302.650	180.349	276.333	164.667	104.016	61.983	
Ser90	**Euclidean**	**observed**	527	301	604	320	194	132	0.1663
		expected	527.959	300.040	589.172	334.827	207.868	118.131	
Ser90	**Voronoi**	**observed**	434	278	387	216	141	93	0.3981
		expected	442.184	269.815	374.490	228.509	145.324	88.675	
Thr30	**Euclidean**	**observed**	194	129	252	123	73	105	**3.343×10^{-8}**
		expected	191.366	131.633	222.174	152.825	105.458	72.541	
Thr30	**Voronoi**	**observed**	164	116	142	91	50	41	0.6058
		expected	165.033	114.966	137.331	95.668	53.635	37.364	
Thr90	**Euclidean**	**observed**	331	171	451	183	146	92	**1.430×10^{-2}**
		expected	339.050	162.949	428.203	205.796	160.745	77.254	
Thr90	**Voronoi**	**observed**	306	156	261	133	108	62	0.7936
		expected	303.947	158.052	259.210	134.789	111.842	58.157	

(B)

dataset	area	type	phobic		philic		polar		p-value
			pP	NonP	pP	NonP	pP	NonP	
Tyr30	Euclidean	observed	76	1 701	108	1 207	38	504	2.350×10^{-5}
		expected	108.556	1 668.443	80.332	1 234.667	33.110	508.889	
Tyr30	Voronoi	observed	106	1 848	65	1 147	39	538	0.4264
		expected	109.628	1 844.371	67.998	1 144.001	32.372	544.627	
Tyr90	Euclidean	observed	134	3 571	260	2 454	96	832	$< 2.2 \times 10^{-16}$
		expected	247.100	3 457.899	181.007	2 532.992	61.891	866.108	
Tyr90	Voronoi	observed	254	3 900	207	2 381	165	992	$< 2.2 \times 10^{-16}$
		expected	329.206	3 824.793	205.100	2 382.899	91.692	1 065.307	
Ser30	Euclidean	observed	378	8 742	443	5 918	146	2 531	1.403×10^{-13}
		expected	485.683	8 634.316	338.753	6 022.246	142.563	2 534.437	
Ser30	Voronoi	observed	294	6 041	286	4 165	103	1 787	2.817×10^{-4}
		expected	341.338	5 993.661	239.825	4 211.174	101.835	1 788.164	
Ser90	Euclidean	observed	527	11 864	604	8 114	194	3 363	$< 2.2 \times 10^{-16}$
		expected	665.615	11 725.384	468.310	8 249.689	191.073	3 365.926	
Ser90	Voronoi	observed	434	8 243	387	5 722	141	2 363	2.297×10^{-3}
		expected	482.780	8 194.219	339.899	5 769.100	139.320	2 364.679	
Thr30	Euclidean	observed	194	4 544	252	2 975	73	1 298	1.020×10^{-11}
		expected	263.391	4 474.608	179.393	3 047.607	76.215	1 294.784	
Thr30	Voronoi	observed	164	3 072	142	2 051	50	810	8.670×10^{-2}
		expected	183.179	3 052.820	124.138	2 068.861	48.681	811.318	
Thr90	Euclidean	observed	331	7 668	451	5 367	146	2 212	$< 2.2 \times 10^{-16}$
		expected	458.922	7 540.077	333.793	5 484.206	135.284	2 222.715	
Thr90	Voronoi	observed	306	5 271	261	3 696	108	1 424	2.053×10^{-2}
		expected	340.183	5 236.816	241.367	3 715.632	93.448	1 438.551	

(C)

dataset	area	type	phobic		philic		polar		p-value
			wP	NonP	wP	NonP	wP	NonP	
Tyr30	Euclidean	observed	79	1 701	54	1 207	33	504	0.1941
		expected	82.582	1 697.417	58.503	1 202.496	24.913	512.086	
Tyr30	Voronoi	observed	87	1 848	63	1 147	37	538	0.1644
		expected	97.270	1 837.729	60.825	1 149.174	28.904	546.095	
Tyr90	Euclidean	observed	182	3 571	171	2 454	67	832	1.359×10^{-3}
		expected	216.608	3 536.391	151.504	2 473.495	51.886	847.113	
Tyr90	Voronoi	observed	201	3 900	173	2 381	77	992	8.107×10^{-4}
		expected	239.455	3 861.544	149.126	2 404.873	62.418	1 006.581	
Ser30	Euclidean	observed	204	8 742	216	5 918	94	2 531	3.909×10^{-6}
		expected	259.714	8 686.285	178.078	5 955.921	76.207	2 548.792	
Ser30	Voronoi	observed	189	6 041	155	4 165	63	1 787	0.2763
		expected	204.484	6 025.515	141.793	4 178.206	60.721	1 789.278	
Ser90	Euclidean	observed	301	11 864	320	8 114	132	3 363	3.405×10^{-8}
		expected	380.187	11 784.812	263.584	8 170.415	109.227	3 385.772	
Ser90	Voronoi	observed	278	8 243	216	5 722	93	2 363	0.3123
		expected	295.703	8 225.296	206.066	5 731.934	85.230	2 370.769	
Thr30	Euclidean	observed	129	4 544	123	2 975	105	1 298	1.013×10^{-14}
		expected	181.846	4 491.153	120.556	2 977.443	54.596	1 348.403	
Thr30	Voronoi	observed	116	3 072	91	2 051	41	810	0.2346
		expected	127.912	3 060.088	85.943	2 056.056	34.144	816.855	
Thr90	Euclidean	observed	171	7 668	183	5 367	92	2 212	1.010×10^{-6}
		expected	222.786	7 616.213	157.732	5 392.267	65.480	2 238.519	
Thr90	Voronoi	observed	156	5 271	133	3 696	62	1 424	2.998×10^{-2}
		expected	177.329	5 249.670	125.114	3 703.885	48.555	1 437.444	

Table 8: Distribution of aa's spatially neighboring the pP-, wP-, and NonP-sites in three groups based on charge ('positive', 'negative', and 'neutral)'. Residues that were not from the list of 20 'canonical 'amino acids, were treated as 'hetatm' (red) and excluded from further analysis. The distribution is presented as the number ('count') of aa's spatially neighboring the site and a relative percentage ('percent'), where sum of all aa'a in particular dataset is 100%. For each amino acid dataset (serine = 'Ser', threonine = 'Thr', and tyrosine = 'Tyr') two subsets (30 and 90) were made, where the number corresponds to the maximum sequence identity shared by the protein chains in the redundancy set. Amino acids spatially neighboring the sites were selected by Euclidean distance and Voronoi diagrams.

dataset	area	type	count					percent		
			positive	negative	neutral	hetatm	sum aa's	positive	negative	neutral
Tyr30	Euclidean	pP	76	17	129	55	222	34.234	7.657	58.108
		wP	33	16	117	34	166	19.879	9.638	70.481
		NonP	544	396	2 472	881	3 412	15.943	11.606	72.450
Tyr30	Voronoi	pP	37	19	154	34	210	17.619	9.047	73.333
		wP	30	23	134	23	187	16.042	12.299	71.657
		NonP	537	341	2 655	401	3 533	15.199	9.651	75.148
Tyr90	Euclidean	pP	187	45	258	152	490	38.163	9.183	52.653
		wP	89	54	277	131	420	21.190	12.857	65.952
		NonP	1 166	758	4 933	1 566	6 857	17.004	11.054	71.941
Tyr90	Voronoi	pP	151	62	413	31	626	24.121	9.904	65.974
		wP	91	64	296	76	451	20.177	14.190	65.631
		NonP	1 149	698	5 426	789	7 273	15.798	9.597	74.604
Ser30	Euclidean	pP	234	120	613	276	967	24.198	12.409	63.391
		wP	99	74	341	141	514	19.260	14.396	66.342
		NonP	2 372	2 117	12 702	3 665	17 191	13.797	12.314	73.887
Ser30	Voronoi	pP	126	90	467	119	683	18.448	13.177	68.374
		wP	71	52	284	60	407	17.444	12.776	69.778
		NonP	1 739	1 432	8 822	1 726	11 993	14.500	11.940	73.559
Ser90	Euclidean	pP	338	152	835	366	1 325	25.509	11.471	63.018
		wP	146	105	502	225	753	19.389	13.944	66.666
		NonP	3 275	2 997	17 069	5 153	23 341	14.031	12.840	73.128
Ser90	Voronoi	pP	195	105	662	157	962	20.270	10.914	68.814
		wP	103	74	410	93	587	17.546	12.606	69.846
		NonP	2 408	2 028	11 892	2 384	16 328	14.747	12.420	72.831
Thr30	Euclidean	pP	154	68	297	126	519	29.672	13.102	57.225
		wP	66	41	210	56	317	20.820	12.933	66.246
		NonP	1 245	1 021	6 551	1 552	8 817	14.120	11.579	74.299
Thr30	Voronoi	pP	74	39	243	51	356	20.786	10.955	68.258
		wP	46	25	177	22	248	18.548	10.080	71.370
		NonP	878	693	4 362	663	5 933	14.798	11.680	73.520
Thr90	Euclidean	pP	297	100	531	245	928	32.004	10.775	57.219
		wP	99	57	290	109	446	22.197	12.780	65.022
		NonP	2 238	1 946	11 063	2 852	15 247	14.678	12.763	72.558
Thr90	Voronoi	pP	146	68	461	113	675	21.629	10.074	68.296
		wP	67	38	246	39	351	19.088	10.826	70.085
		NonP	1 568	1 320	7 503	1 251	10 391	15.089	12.703	72.206

Table 9: The distribution of aa's spatially neighboring (A) pP- and wP-sites, (B) pP- and NonP-sites, and (C) wP- and NonP-sites in three groups based on charge ('positive', 'negative', and 'neutral)'. For each amino acid dataset (serine = 'Ser', threonine = 'Thr', and tyrosine = 'Tyr') two subsets (30 and 90) were made, where the number corresponds to the maximum sequence identity shared by the protein chains in the redundancy set. Observed and chi-squared test expected frequencies are presented. When the difference between distributions was significant at the 0.05 level, p-value was marked in bold.

(A)

dataset	area	type	positive		negative		neutral		p-value
			pP	wP	pP	wP	pP	wP	
Tyr30	**Euclidean**	**observed**	76	33	17	16	129	117	**7.834×10^{-3}**
		expected	62.365	46.634	18.881	14.118	140.752	105.247	
Tyr30	**Voronoi**	**observed**	37	30	19	23	154	134	0.5564
		expected	35.440	31.559	22.216	19.783	152.342	135.657	
Tyr90	**Euclidean**	**observed**	187	89	45	54	258	277	**1.774×10^{-7}**
		expected	148.615	127.384	53.307	45.692	288.076	246.923	
Tyr90	**Voronoi**	**observed**	151	91	62	64	413	296	0.0513
		expected	140.661	101.338	73.236	52.763	412.102	296.897	
Ser30	**Euclidean**	**observed**	234	99	120	74	613	341	7.789×10^{-2}
		expected	217.428	115.571	126.669	67.330	622.902	331.097	
Ser30	**Voronoi**	**observed**	126	71	90	52	467	284	0.8841
		expected	123.441	73.558	88.977	53.022	470.580	280.419	
Ser90	**Euclidean**	**observed**	338	146	152	105	835	502	**3.911×10^{-3}**
		expected	308.614	175.385	163.871	93.128	852.514	484.485	
Ser90	**Voronoi**	**observed**	195	103	105	74	662	410	0.3066
		expected	185.071	112.928	111.167	67.832	665.761	406.238	
Thr30	**Euclidean**	**observed**	154	66	68	41	297	210	**1.423×10^{-2}**
		expected	136.578	83.421	67.668	41.331	314.752	192.247	
Thr30	**Voronoi**	**observed**	74	46	39	25	243	177	0.7126
		expected	70.728	49.271	37.721	26.278	247.549	172.450	
Thr90	**Euclidean**	**observed**	297	99	100	57	531	290	**8.332×10^{-4}**
		expected	267.458	128.541	106.037	50.962	554.503	266.496	
Thr90	**Voronoi**	**observed**	146	67	68	38	461	246	0.6212
		expected	140.131	72.868	69.736	36.263	465.131	241.868	

(B)

dataset	area	type	positive		negative		neutral		p-value
			pP	NonP	pP	NonP	pP	NonP	
Tyr30	Euclidean	observed	76	544	17	396	129	2 472	**1.597×10^{-11}**
		expected	37.875	582.124	25.230	387.769	158.894	2 442.105	
Tyr30	Voronoi	observed	37	537	19	341	154	2 655	0.6316
		expected	32.204	541.795	20.197	339.802	157.598	2 651.401	
Tyr90	Euclidean	observed	187	1 166	45	758	258	4 933	**$< 2.2 \times 10^{-16}$**
		expected	90.236	1 262.763	53.555	749.444	346.208	4 844.792	
Tyr90	Voronoi	observed	151	1 149	62	698	413	5 426	**2.872×10^{-7}**
		expected	103.025	1 196.974	60.230	699.769	462.743	5 376.256	
Ser30	Euclidean	observed	234	2 372	120	2 117	613	12 702	**$< 2.2 \times 10^{-16}$**
		expected	138.781	2 467.218	119.130	2 117.869	709.087	12 605.912	
Ser30	Voronoi	observed	126	1 739	90	1 432	467	8 822	**6.607×10^{-3}**
		expected	100.488	1 764.511	82.007	1 439.992	500.503	8 788.496	
Ser90	Euclidean	observed	338	3 275	152	2 997	835	17 069	**$< 2.2 \times 10^{-16}$**
		expected	194.081	3 418.918	169.156	2 979.843	961.761	16 942.238	
Ser90	Voronoi	observed	195	2 408	105	2 028	662	11 892	**1.595×10^{-5}**
		expected	144.828	2 458.171	118.678	2 014.321	698.493	11 855.506	
Thr30	Euclidean	observed	154	1 245	68	1 021	297	6 551	**$< 2.2 \times 10^{-16}$**
		expected	77.772	1 321.227	60.538	1 028.461	380.688	6 467.311	
Thr30	Voronoi	observed	74	878	39	693	243	4 362	**9.200×10^{-3}**
		expected	53.889	898.110	41.436	690.563	260.674	4 344.325	
Thr90	Euclidean	observed	297	2 238	100	1 946	531	11 063	**$< 2.2 \times 10^{-16}$**
		expected	145.439	2 389.560	117.384	1 928.615	665.176	10 928.823	
Thr90	Voronoi	observed	146	1 568	68	1 320	461	7 503	**1.409×10^{-5}**
		expected	104.550	1 609.450	84.664	1 303.335	485.785	7 478.214	

(C)

dataset	area	type	positive		negative		neutral		p-value
			wP	NonP	wP	NonP	wP	NonP	
Tyr30	Euclidean	observed	33	544	16	396	117	2 472	0.3434
		expected	26.769	550.230	19.114	392.885	120.115	2 468.884	
Tyr30	Voronoi	observed	30	537	23	341	134	2 655	0.4396
		expected	28.502	538.497	18.297	345.702	140.199	2 648.800	
Tyr90	Euclidean	observed	89	1 166	54	758	277	4 933	**2.790×10^{-2}**
		expected	72.433	1 182.566	46.865	765.134	300.700	4 909.299	
Tyr90	Voronoi	observed	91	1 149	64	698	296	5 426	**8.392×10^{-5}**
		expected	72.402	1 167.597	44.492	717.507	334.104	5 387.895	
Ser30	Euclidean	observed	99	2 372	74	2 117	341	12 702	**3.796×10^{-4}**
		expected	71.494	2 399.505	65.128	2 185.871	377.377	12 665.623	
Ser30	Voronoi	observed	71	1 739	52	1 432	284	8 822	0.1888
		expected	59.408	1 750.591	48.708	1 435.291	298.882	8 807.117	
Ser90	Euclidean	observed	146	3 275	105	2 997	502	17 069	**5.498×10^{-5}**
		expected	106.915	3 314.084	96.945	3 005.054	549.139	17 021.860	
Ser90	Voronoi	observed	103	2 408	74	2 028	410	11 892	0.1572
		expected	87.139	2 423.860	72.945	2 029.054	426.915	11 875.084	
Thr30	Euclidean	observed	66	1 245	41	1 021	210	6 551	**1.718×10^{-3}**
		expected	45.498	1 265.501	36.857	1 025.142	234.643	6 526.356	
Thr30	Voronoi	observed	46	878	25	693	177	4 362	0.2330
		expected	37.073	886.926	28.808	689.191	182.118	4 356.881	
Thr90	Euclidean	observed	99	2 238	57	1 946	290	11 063	**4.888×10^{-5}**
		expected	66.418	2 270.581	56.925	1 946.074	322.655	11 030.344	
Thr90	Voronoi	observed	67	1 568	38	1 320	246	7 503	9.422×10^{-2}
		expected	53.424	1 581.575	44.373	1 313.626	253.202	7 495.797	

Table 10: The average conservation level of the sites and aa's neighboring the sites selected by Euclidean distance and Voronoi diagrams. For each amino acid dataset (serine = 'Ser', threonine = 'Thr', and tyrosine = 'Tyr') two subsets (30 and 90) were made, where the number corresponds to the maximum sequence identity shared by the protein chains in the redundancy set.

dataset		sites	neighboring aa's	
			Euclidean	Voronoi
Tyr30	pP	7.090	6.328	7.214
	wP	6.500	6.720	6.881
	NonP	5.568	6.002	5.965
Tyr90	pP	6.634	6.271	6.017
	wP	6.222	6.768	6.312
	NonP	5.415	6.255	6.200
Ser30	pP	5.555	6.408	6.005
	wP	6.280	6.157	6.020
	NonP	5.052	6.255	4.969
Ser90	pP	6.131	6.577	5.882
	wP	6.171	6.271	6.031
	NonP	5.087	5.988	5.167
Thr30	pP	6.583	5.578	5.354
	wP	5.714	5.786	5.664
	NonP	4.670	7.540	6.082
Thr90	pP	6.945	6.880	6.235
	wP	6.350	5.895	5.982
	NonP	5.272	6.421	5.813

Table 11: Results of multiple ANOVA analysis of the sites conservation. The means of average distance between sites and aa's neighboring the sites selected by Euclidean distance and Voronoi diagrams were calculated and compared between datasets (datasets1-datasets2). Mean difference between datasets, standard error ('Std. Error'), significance ('Sig.'), and 95% confidence interval are presented. ANOVA analysis was made for (1A) Tyr30, Ser30, and Thr30 sites, (1B) Tyr90, Ser90, and Thr90 sites, (2A) aa's spatially neighboring tyrosine sites, (2B) aa's spatially neighboring serine sites, and (2C) aa's spatially neighboring threonine sites. For each amino acid dataset (serine = 'Ser', threonine = 'Thr', and tyrosine = 'Tyr') two subsets (30 and 90) were made, where the number corresponds to the maximum sequence identity shared by the protein chains in the redundancy set.

(1A)

Multiple Comparisons (Tukey HSD)																					
dataset1	dataset2	Mean Difference (1-2)	Std. Error	Sig.	95% Confidence Interval		dataset1	dataset2	Mean Difference (1-2)	Std. Error	Sig.	95% Confidence Interval		dataset1	dataset2	Mean Difference (1-2)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound						Lower Bound	Upper Bound						Lower Bound	Upper Bound
Tyr30pP	Ser30NonP	2.038	0.809	0.516	-0.788	4.865	Tyr30wP	Ser30NonP	1.447	0.945	0.988	-1.855	4.750	Tyr30NonP	Ser30NonP	0.516	0.240	0.788	-0.323	1.356
	Ser30pP	1.535	0.912	0.970	-1.651	4.722		Ser30pP	0.944	1.035	1.000	-2.671	4.559		Ser30pP	0.013	0.484	1.000	-1.680	1.707
	Ser30wP	0.811	0.958	1.000	-2.535	4.157		Ser30wP	0.220	1.075	1.000	-3.537	3.977		Ser30wP	-0.711	0.566	0.999	-2.689	1.267
	Thr30NonP	2.420	0.850	0.285	-0.550	5.390		Thr30NonP	1.829	0.980	0.924	-1.596	5.255		Thr30NonP	0.898	0.354	0.506	-0.341	2.137
	Thr30pP	0.508	1.105	1.000	-3.353	4.368		Thr30pP	-0.083	1.208	1.000	-4.305	4.138		Thr30pP	-1.014	0.790	0.998	-3.775	1.746
	Thr30wP	1.377	1.067	0.998	-2.350	5.103		Thr30wP	0.786	1.173	1.000	-3.313	4.885		Thr30wP	-0.145	0.735	1.000	-2.715	2.424
	Tyr30NonP	1.522	0.823	0.930	-1.353	4.397		Tyr30NonP	0.931	0.957	1.000	-2.413	4.275		Tyr30pP	-1.522	0.823	0.930	-4.397	1.353
	Tyr30wP	0.591	1.230	1.000	-3.707	4.888		Tyr30pP	-0.591	1.230	1.000	-4.888	3.707		Tyr30wP	-0.931	0.957	1.000	-4.275	2.413
Ser30pP	Ser30NonP	0.503	0.460	1.000	-1.106	2.112	Ser30wP	Ser30NonP	1.228	0.545	0.721	-0.679	3.134	Ser30NonP	Ser30pP	-0.503	0.460	1.000	-2.112	1.106
	Ser30wP	-0.724	0.689	1.000	-3.132	1.683		Ser30pP	0.724	0.689	1.000	-1.683	3.132		Ser30wP	-1.227	0.545	0.721	-3.134	0.679
	Thr30NonP	0.885	0.529	0.972	-0.964	2.734		Thr30NonP	1.609	0.605	0.410	-0.504	3.722		Thr30NonP	0.382	0.321	0.999	-0.739	1.503
	Thr30pP	-1.028	0.882	1.000	-4.111	2.055		Thr30pP	-0.303	0.930	1.000	-3.551	2.945		Thr30pP	-1.531	0.775	0.882	-4.240	1.179
	Thr30wP	-0.159	0.834	1.000	-3.072	2.754		Thr30wP	0.566	0.884	1.000	-2.522	3.653		Thr30wP	-0.662	0.720	1.000	-3.176	1.853
	Tyr30NonP	-0.013	0.484	1.000	-1.707	1.680		Tyr30NonP	0.711	0.566	0.999	-1.267	2.689		Tyr30NonP	-0.516	0.240	0.788	-1.356	0.323
	Tyr30pP	-1.535	0.912	0.970	-4.722	1.651		Tyr30pP	-0.811	0.958	1.000	-4.157	2.535		Tyr30pP	-2.038	0.809	0.516	-4.865	0.788
	Tyr30wP	-0.944	1.035	1.000	-4.559	2.671		Tyr30wP	-0.220	1.075	1.000	-3.977	3.537		Tyr30wP	-1.447	0.945	0.988	-4.750	1.855
Thr30pP	Ser30NonP	1.531	0.775	0.882	-1.179	4.240	Thr30wP	Ser30NonP	0.662	0.720	1.000	-1.853	3.176	Thr30NonP	Ser30NonP	-0.382	0.321	0.999	-1.503	0.739
	Ser30pP	1.028	0.882	1.000	-2.055	4.111		Ser30pP	0.159	0.834	1.000	-2.754	3.072		Ser30pP	-0.885	0.529	0.972	-2.734	0.964
	Ser30wP	0.303	0.930	1.000	-2.945	3.551		Ser30wP	-0.566	0.884	1.000	-3.653	2.522		Ser30wP	-1.609	0.605	0.410	-3.722	0.504
	Thr30NonP	1.913	0.818	0.656	-0.946	4.771		Thr30NonP	1.044	0.765	0.997	-1.631	3.718		Thr30pP	-1.913	0.818	0.656	-4.771	0.946
	Thr30wP	0.869	1.041	1.000	-2.769	4.507		Thr30pP	-0.869	1.041	1.000	-4.507	2.769		Thr30wP	-1.044	0.765	0.997	-3.718	1.631
	Tyr30NonP	1.014	0.790	0.998	-1.746	3.775		Tyr30NonP	0.145	0.735	1.000	-2.424	2.715		Tyr30NonP	-0.898	0.354	0.506	-2.137	0.341
	Tyr30pP	-0.508	1.105	1.000	-4.368	3.353		Tyr30pP	-1.377	1.067	0.998	-5.103	2.350		Tyr30pP	-2.420	0.850	0.285	-5.390	0.550
	Tyr30wP	0.083	1.208	1.000	-4.138	4.305		Tyr30wP	-0.786	1.173	1.000	-4.885	3.313		Tyr30wP	-1.829	0.980	0.924	-5.255	1.596

(1B)

Multiple Comparisons (Tukey HSD)																					
dataset1	dataset2	Mean Difference (1-2)	Std. Error	Sig.	95% Confidence Interval		dataset1	dataset2	Mean Difference (1-2)	Std. Error	Sig.	95% Confidence Interval		dataset1	dataset2	Mean Difference (1-2)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound						Lower Bound	Upper Bound						Lower Bound	Upper Bound
Tyr90pP	Ser90NonP	1.547*	0.425	0.031	0.063	3.031	Tyr90wP	Ser90NonP	1.135	0.518	0.762	-0.677	2.947	Tyr90NonP	Ser90NonP	0.328	0.163	0.865	-0.242	0.899
	Ser90pP	0.503	0.534	1.000	-1.365	2.371		Ser90pP	0.091	0.612	1.000	-2.047	2.229		Ser90pP	-0.715	0.363	0.884	-1.985	0.554
	Ser90wP	0.463	0.609	1.000	-1.666	2.591		Ser90wP	0.051	0.678	1.000	-2.318	2.420		Ser90wP	-0.756	0.466	0.979	-2.385	0.873
	Thr90NonP	1.362	0.433	0.140	-0.151	2.875		Thr90NonP	0.950	0.525	0.942	-0.886	2.786		Thr90NonP	0.143	0.183	1.000	-0.499	0.785
	Thr90pP	-0.312	0.600	1.000	-2.409	1.785		Thr90pP	-0.724	0.670	1.000	-3.065	1.617		Thr90pP	-1.530	0.454	0.074	-3.118	0.057
	Thr90wP	0.284	0.722	1.000	-2.238	2.807		Thr90wP	-0.128	0.781	1.000	-2.856	2.601		Thr90wP	-0.934	0.606	0.988	-3.052	1.184
	Tyr90NonP	1.218	0.433	0.308	-0.297	2.734		Tyr90NonP	0.807	0.526	0.988	-1.031	2.644		Tyr90pP	-1.218	0.433	0.308	-2.734	0.297
	Tyr90wP	0.412	0.656	1.000	-1.880	2.704		Tyr90pP	-0.412	0.656	1.000	-2.704	1.880		Tyr90wP	-0.807	0.526	0.988	-2.644	1.031
Ser90pP	Ser90NonP	1.084	0.458	0.634	-0.516	2.684	Ser90wP	Ser90NonP	1.044	0.352	0.222	-0.188	2.276	Ser90NonP	Ser90pP	-1.044	0.352	0.222	-2.276	0.188
	Ser90wP	0.040	0.561	1.000	-1.921	2.001		Ser90pP	-0.040	0.561	1.000	-2.001	1.921		Ser90wP	-1.084	0.458	0.634	-2.684	0.516
	Thr90NonP	0.899	0.465	0.901	-0.728	2.526		Thr90NonP	0.859	0.362	0.633	-0.408	2.125		Thr90NonP	-0.185	0.161	1.000	-0.750	0.380
	Thr90pP	-0.775	0.624	0.999	-2.955	1.406		Thr90pP	-0.815	0.551	0.992	-2.742	1.112		Thr90pP	-1.859*	0.446	0.004	-3.417	-0.300
	Thr90wP	-0.179	0.742	1.000	-2.771	2.414		Thr90wP	-0.219	0.682	1.000	-2.602	2.164		Thr90wP	-1.263	0.600	0.814	-3.359	0.833
	Tyr90NonP	0.756	0.466	0.979	-0.873	2.385		Tyr90NonP	0.715	0.363	0.884	-0.554	1.985		Tyr90NonP	-0.328	0.163	0.865	-0.899	0.242
	Tyr90pP	-0.463	0.609	1.000	-2.591	1.666		Tyr90pP	-0.503	0.534	1.000	-2.371	1.365		Tyr90pP	-1.547*	0.425	0.031	-3.031	-0.063
	Tyr90wP	-0.051	0.678	1.000	-2.420	2.318		Tyr90wP	-0.091	0.612	1.000	-2.229	2.047		Tyr90wP	-1.135	0.518	0.762	-2.947	0.677
Thr90pP	Ser90NonP	1.859*	0.446	0.004	0.300	3.417	Thr90wP	Ser90NonP	1.263	0.600	0.814	-0.833	3.359	Thr90NonP	Ser90NonP	0.185	0.161	1.000	-0.380	0.750
	Ser90pP	0.815	0.551	0.992	-1.112	2.742		Ser90pP	0.219	0.682	1.000	-2.164	2.602		Ser90pP	-0.859	0.362	0.633	-2.125	0.408
	Ser90wP	0.775	0.624	0.999	-1.406	2.955		Ser90wP	0.179	0.742	1.000	-2.414	2.771		Ser90wP	-0.899	0.465	0.901	-2.526	0.728
	Thr90NonP	1.673*	0.454	0.026	0.088	3.259		Thr90NonP	1.077	0.606	0.950	-1.039	3.194		Thr90pP	-1.673*	0.454	0.026	-3.259	-0.088
	Thr90wP	0.596	0.735	1.000	-1.971	3.163		Thr90pP	-0.596	0.735	1.000	-3.163	1.971		Thr90wP	-1.077	0.606	0.950	-3.194	1.039
	Tyr90NonP	1.530	0.454	0.074	-0.057	3.118		Tyr90NonP	0.934	0.606	0.988	-1.184	3.052		Tyr90NonP	-0.143	0.183	1.000	-0.785	0.499
	Tyr90pP	0.312	0.600	1.000	-1.785	2.409		Tyr90pP	-0.284	0.722	1.000	-2.807	2.238		Tyr90pP	-1.362	0.433	0.140	-2.875	0.151
	Tyr90wP	0.724	0.670	1.000	-1.617	3.065		Tyr90wP	0.128	0.781	1.000	-2.601	2.856		Tyr90wP	-0.950	0.525	0.942	-2.786	0.886

(2A)

Multiple Comparisons (Tukey HSD)																					
Method	dataset1	dataset2	Mean Difference (1-2)	Std. Error	Sig.	95% Confidence Interval		dataset1	dataset2	Mean Difference (1-2)	Std. Error	Sig.	95% Confidence Interval		dataset1	dataset2	Mean Difference (1-2)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound						Lower Bound	Upper Bound						Lower Bound	Upper Bound
Euclidean	Tyr30pP	Ser30NonP	0.072	0.511	1.000	-1.715	1.860	Tyr30wP	Ser30NonP	0.464	0.540	1.000	-1.424	2.353	Tyr30NonP	Ser30NonP	-0.253	0.303	1.000	-1.314	0.807
		Ser30pP	-0.079	0.565	1.000	-2.057	1.897		Ser30pP	0.311	0.591	1.000	-1.757	2.381		Ser30pP	-0.405	0.387	1.000	-1.762	0.950
		Ser30wP	0.170	0.573	1.000	-1.835	2.177		Ser30wP	0.562	0.599	1.000	-1.533	2.659		Ser30wP	-0.154	0.399	1.000	-1.552	1.243
		Thr30NonP	-1.212	0.522	0.669	-3.040	0.615		Thr30NonP	-0.820	0.550	0.991	-2.747	1.105		Thr30NonP	-1.538*	0.322	0.000	-2.664	-0.411
		Thr30pP	0.750	0.663	1.000	-1.569	3.070		Thr30pP	1.142	0.685	0.973	-1.256	3.540		Thr30pP	0.424	0.520	1.000	-1.394	2.244
		Thr30wP	0.541	0.630	1.000	-1.663	2.747		Thr30wP	0.933	0.654	0.995	-1.354	3.221		Thr30wP	0.215	0.477	1.000	-1.454	1.886
		Tyr30NonP	0.325	0.561	1.000	-1.636	2.287		Tyr30NonP	0.717	0.587	0.999	-1.336	2.771		Tyr30pP	-0.325	0.561	1.000	-2.287	1.636
		Tyr30wP	-0.391	0.717	1.000	-2.900	2.116		Tyr30pP	0.391	0.717	1.000	-2.116	2.900		Tyr30wP	-0.717	0.587	0.999	-2.771	1.336
	Tyr90pP	Ser90NonP	0.283	0.254	1.000	-0.605	1.172	Tyr90wP	Ser90NonP	0.779	0.299	0.449	-0.266	1.825	Tyr90NonP	Ser90NonP	0.266	0.146	0.938	-0.245	0.778
		Ser90pP	-0.305	0.315	1.000	-1.408	0.796		Ser90pP	0.190	0.352	1.000	-1.042	1.423		Ser90pP	-0.322	0.237	0.997	-1.151	0.506
		Ser90wP	-7.000×10^{-5}	0.343	1.000	-1.201	1.201		Ser90wP	0.496	0.378	0.998	-0.825	1.818		Ser90wP	-0.016	0.273	1.000	-0.973	0.940
		Thr90NonP	-0.149	0.253	1.000	-1.036	0.738		Thr90NonP	0.347	0.298	1.000	-0.697	1.391		Thr90NonP	-0.165	0.145	1.000	-0.674	0.342
		Thr90pP	-0.608	0.338	0.945	-1.793	0.575		Thr90pP	-0.112	0.373	1.000	-1.418	1.194		Thr90pP	-0.625	0.267	0.655	-1.560	0.309
		Thr90wP	0.376	0.413	1.000	-1.067	1.820		Thr90wP	0.872	0.442	0.881	-0.672	2.418		Thr90wP	0.359	0.356	1.000	-0.888	1.607
		Tyr90NonP	0.016	0.261	1.000	-0.898	0.932		Tyr90NonP	0.513	0.305	0.970	-0.555	1.581		Tyr90pP	-0.016	0.261	1.000	-0.932	0.898
		Tyr90wP	-0.496	0.369	0.997	-1.788	0.795		Tyr90pP	0.496	0.369	0.997	-0.795	1.788		Tyr90wP	-0.513	0.305	0.970	-1.581	0.555
Voronoi	Tyr30pP	Ser30NonP	2.245*	0.347	0.000	1.031	3.458	Tyr30wP	Ser30NonP	1.911*	0.387	0.000	0.558	3.264	Tyr30NonP	Ser30NonP	0.995*	0.100	0.000	0.643	1.347
		Ser30pP	1.209	0.399	0.188	-0.186	2.604		Ser30pP	0.875	0.434	0.863	-0.642	2.394		Ser30pP	-0.040	0.221	1.000	-0.814	0.734
		Ser30wP	1.194	0.405	0.230	-0.222	2.611		Ser30wP	0.861	0.440	0.890	-0.677	2.399		Ser30wP	-0.054	0.232	1.000	-0.868	0.758
		Thr30NonP	1.131	0.363	0.151	-0.137	2.401		Thr30NonP	0.798	0.402	0.876	-0.605	2.202		Thr30NonP	-0.117	0.147	1.000	-0.631	0.396
		Thr30pP	1.860*	0.473	0.011	0.205	3.515		Thr30pP	1.526	0.504	0.187	-0.233	3.287		Thr30pP	0.610	0.337	0.942	-0.568	1.790
		Thr30wP	1.550	0.449	0.057	-0.018	3.118		Thr30wP	1.216	0.480	0.506	-0.462	2.895		Thr30wP	0.300	0.301	1.000	-0.753	1.355
		Tyr30NonP	1.249*	0.353	0.043	0.015	2.483		Tyr30NonP	0.916	0.392	0.659	-0.455	2.287		Tyr30pP	-1.249*	0.353	0.043	-2.483	-0.015
		Tyr30wP	0.333	0.514	1.000	-1.463	2.130		Tyr30pP	-0.333	0.514	1.000	-2.130	1.463		Tyr30wP	-0.916	0.392	0.659	-2.287	0.455
	Tyr90pP	Ser90NonP	0.849*	0.176	0.000	0.233	1.465	Tyr90wP	Ser90NonP	1.144*	0.212	0.000	0.401	1.887	Tyr90NonP	Ser90NonP	1.032*	0.067	0.000	0.795	1.269
		Ser90pP	0.134	0.230	1.000	-0.669	0.938		Ser90pP	0.429	0.259	0.974	-0.475	1.334		Ser90pP	0.317	0.162	0.891	-0.250	0.886
		Ser90wP	-0.014	0.251	1.000	-0.891	0.862		Ser90wP	0.280	0.277	1.000	-0.689	1.251		Ser90wP	0.169	0.191	1.000	-0.498	0.836
		Thr90NonP	0.203	0.179	1.000	-0.423	0.830		Thr90NonP	0.498	0.215	0.673	-0.253	1.251		Thr90NonP	0.386*	0.076	0.000	0.120	0.652
		Thr90pP	-0.218	0.247	1.000	-1.082	0.645		Thr90pP	0.076	0.274	1.000	-0.882	1.035		Thr90pP	-0.035	0.186	1.000	-0.686	0.615
		Thr90wP	0.034	0.302	1.000	-1.021	1.089		Thr90wP	0.329	0.324	1.000	-0.804	1.463		Thr90wP	0.217	0.254	1.000	-0.671	1.107
		Tyr90NonP	-0.183	0.179	1.000	-0.811	0.445		Tyr90NonP	0.111	0.215	1.000	-0.641	0.865		Tyr90pP	0.183	0.179	1.000	-0.445	0.811
		Tyr90wP	-0.295	0.270	1.000	-1.238	0.648		Tyr90pP	0.295	0.270	1.000	-0.648	1.238		Tyr90wP	-0.111	0.215	1.000	-0.865	0.641

(2B)

	Multiple Comparisons (Tukey HSD)																			
Method	dataset1	dataset2	Mean Difference (1-2)	Std. Error	Sig.	95% Confidence Interval		dataset1	dataset2	Mean Difference (1-2)	Std. Error	Sig.	95% Confidence Interval		dataset1	dataset2	Mean Difference (1-2)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound						Lower Bound	Upper Bound						Lower Bound	Upper Bound
Euclidean	Ser30pP	Ser30NonP	0.152	0.311	1.000	-0.938	1.243	Ser30wP	Ser30NonP	-0.098	0.326	1.000	-1.240	1.043	Ser30NonP	Ser30pP	-0.152	0.311	1.000	-1.243	0.938
		Ser30wP	0.250	0.406	1.000	-1.169	1.671		Ser30pP	-0.250	0.406	1.000	-1.671	1.169		Ser30wP	0.098	0.326	1.000	-1.043	1.240
		Thr30NonP	-1.132	0.330	0.061	-2.286	0.021		Thr30NonP	-1.383*	0.343	0.008	-2.586	-0.180		Thr30NonP	-1.285*	0.224	0.000	-2.071	-0.498
		Thr30pP	0.830	0.525	0.984	-1.006	2.667		Thr30pP	0.579	0.534	1.000	-1.288	2.446		Thr30pP	0.677	0.466	0.993	-0.953	2.308
		Thr30wP	0.621	0.483	0.998	-1.067	2.311		Thr30wP	0.370	0.492	1.000	-1.352	2.093		Thr30wP	0.469	0.418	1.000	-0.994	1.932
		Tyr30NonP	0.405	0.387	1.000	-0.950	1.762		Tyr30NonP	0.154	0.399	1.000	-1.243	1.552		Tyr30NonP	0.253	0.303	1.000	-0.807	1.314
		Tyr30pP	0.079	0.565	1.000	-1.897	2.057		Tyr30pP	-0.170	0.573	1.000	-2.177	1.835		Tyr30pP	-0.072	0.511	1.000	-1.860	1.715
		Tyr30wP	-0.311	0.591	1.000	-2.381	1.757		Tyr30wP	-0.562	0.599	1.000	-2.659	1.533		Tyr30wP	-0.464	0.540	1.000	-2.353	1.424
	Ser90pP	Ser90NonP	0.589	0.228	0.473	-0.210	1.389	Ser90wP	Ser90NonP	0.283	0.266	1.000	-0.648	1.215	Ser90NonP	Ser90pP	-0.589	0.228	0.473	-1.389	0.210
		Ser90wP	0.305	0.325	1.000	-0.831	1.443		Ser90pP	-0.305	0.325	1.000	-1.443	0.831		Ser90wP	-0.283	0.266	1.000	-1.215	0.648
		Thr90NonP	0.156	0.228	1.000	-0.641	0.954		Thr90NonP	-0.149	0.266	1.000	-1.079	0.781		Thr90NonP	-0.432	0.131	0.093	-0.892	0.026
		Thr90pP	-0.302	0.320	1.000	-1.422	0.816		Thr90pP	-0.608	0.348	0.957	-1.825	0.608		Thr90pP	-0.892	0.260	0.062	-1.801	0.017
		Thr90wP	0.682	0.397	0.964	-0.708	2.073		Thr90wP	0.376	0.420	1.000	-1.094	1.847		Thr90wP	0.092	0.351	1.000	-1.135	1.321
		Tyr90NonP	0.322	0.237	0.997	-0.506	1.151		Tyr90NonP	0.016	0.273	1.000	-0.940	0.973		Tyr90NonP	-0.266	0.146	0.938	-0.778	0.245
		Tyr90pP	0.305	0.315	1.000	-0.796	1.408		Tyr90pP	7.000×10^{-5}	0.343	1.000	-1.201	1.201		Tyr90pP	-0.283	0.254	1.000	-1.172	0.605
		Tyr90wP	-0.190	0.352	1.000	-1.423	1.042		Tyr90wP	-0.496	0.378	0.998	-1.818	0.825		Tyr90wP	-0.779	0.299	0.449	-1.825	0.266
Voronoi	Ser30pP	Ser30NonP	1.035*	0.212	0.000	0.294	1.777	Ser30wP	Ser30NonP	1.050*	0.223	0.000	0.269	1.832	Ser30NonP	Ser30pP	-1.035*	0.212	0.000	-1.777	-0.294
		Ser30wP	-0.014	0.298	1.000	-1.057	1.027		Ser30pP	0.014	0.298	1.000	-1.027	1.057		Ser30wP	-1.050*	0.223	0.000	-1.832	-0.269
		Thr30NonP	-0.077	0.237	1.000	-0.907	0.753		Thr30NonP	-0.062	0.248	1.000	-0.929	0.804		Thr30NonP	-1.113*	0.132	0.000	-1.575	-0.650
		Thr30pP	0.651	0.385	0.969	-0.697	1.999		Thr30pP	0.665	0.392	0.967	-0.704	2.036		Thr30pP	-0.384	0.331	1.000	-1.543	0.773
		Thr30wP	0.340	0.355	1.000	-0.899	1.581		Thr30wP	0.355	0.362	1.000	-0.908	1.620		Thr30wP	-0.694	0.295	0.642	-1.725	0.335
		Tyr30NonP	0.040	0.221	1.000	-0.734	0.814		Tyr30NonP	0.054	0.232	1.000	-0.758	0.868		Tyr30NonP	-0.995*	0.100	0.000	-1.347	-0.643
		Tyr30pP	-1.209	0.399	0.188	-2.604	0.186		Tyr30pP	-1.194	0.405	0.230	-2.611	0.222		Tyr30pP	-2.245*	0.347	0.000	-3.458	-1.031
		Tyr30wP	-0.875	0.434	0.863	-2.394	0.642		Tyr30wP	-0.861	0.440	0.890	-2.399	0.677		Tyr30wP	-1.911*	0.387	0.000	-3.264	-0.558
	Ser90pP	Ser90NonP	0.714*	0.158	0.001	0.160	1.269	Ser90wP	Ser90NonP	0.863*	0.187	0.001	0.207	1.519	Ser90NonP	Ser90pP	-0.714*	0.158	0.001	-1.269	-0.160
		Ser90wP	-0.148	0.239	1.000	-0.983	0.686		Ser90pP	0.148	0.239	1.000	-0.686	0.983		Ser90wP	-0.863*	0.187	0.001	-1.519	-0.207
		Thr90NonP	0.069	0.162	1.000	-0.498	0.636		Thr90NonP	0.217	0.190	1.000	-0.449	0.884		Thr90NonP	-0.645*	0.067	0.000	-0.880	-0.411
		Thr90pP	-0.352	0.235	0.991	-1.174	0.468		Thr90pP	-0.204	0.255	1.000	-1.097	0.689		Thr90pP	-1.067*	0.182	0.000	-1.706	-0.429
		Thr90wP	-0.100	0.292	1.000	-1.121	0.920		Thr90wP	0.048	0.309	1.000	-1.031	1.128		Thr90wP	-0.815	0.252	0.109	-1.695	0.065
		Tyr90NonP	-0.317	0.162	0.891	-0.886	0.250		Tyr90NonP	-0.169	0.191	1.000	-0.836	0.498		Tyr90NonP	-1.032*	0.067	0.000	-1.269	-0.795
		Tyr90pP	-0.134	0.230	1.000	-0.938	0.669		Tyr90pP	0.014	0.251	1.000	-0.862	0.891		Tyr90pP	-0.849*	0.176	0.000	-1.465	-0.233
		Tyr90wP	-0.429	0.259	0.974	-1.334	0.475		Tyr90wP	-0.280	0.277	1.000	-1.251	0.689		Tyr90wP	-1.144*	0.212	0.000	-1.887	-0.401

(2C)

Multiple Comparisons (Tukey HSD)																					
Method	dataset1	dataset2	Mean Difference (1-2)	Std. Error	Sig.	95% Confidence Interval		dataset1	dataset2	Mean Difference (1-2)	Std. Error	Sig.	95% Confidence Interval		dataset1	dataset2	Mean Difference (1-2)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound						Lower Bound	Upper Bound						Lower Bound	Upper Bound
Euclidean	Thr30pP	Ser30NonP	-0.677	0.466	0.993	-2.308	0.953	Thr30wP	Ser30NonP	-0.469	0.418	1.000	-1.932	0.994	Thr30NonP	Ser30NonP	1.285*	0.224	0.000	0.498	2.071
		Ser30pP	-0.830	0.525	0.984	-2.667	1.006		Ser30pP	-0.621	0.483	0.998	-2.311	1.067		Ser30pP	1.132	0.330	0.061	-0.021	2.286
		Ser30wP	-0.579	0.534	1.000	-2.446	1.288		Ser30wP	-0.370	0.492	1.000	-2.093	1.352		Ser30wP	1.383*	0.343	0.008	0.180	2.586
		Thr30NonP	-1.962*	0.478	0.006	-3.637	-0.288		Thr30NonP	-1.754*	0.432	0.007	-3.265	-0.242		Thr30pP	1.962*	0.478	0.006	0.288	3.637
		Thr30wP	-0.208	0.594	1.000	-2.288	1.871		Thr30pP	0.208	0.594	1.000	-1.871	2.288		Thr30wP	1.754*	0.432	0.007	0.242	3.265
		Tyr30NonP	-0.424	0.520	1.000	-2.244	1.394		Tyr30NonP	-0.215	0.477	1.000	-1.886	1.454		Tyr30NonP	1.538*	0.322	0.000	0.411	2.664
		Tyr30pP	-0.750	0.663	1.000	-3.070	1.569		Tyr30pP	-0.541	0.630	1.000	-2.747	1.663		Tyr30pP	1.212	0.522	0.669	-0.615	3.040
		Tyr30wP	-1.142	0.685	0.973	-3.540	1.256		Tyr30wP	-0.933	0.654	0.995	-3.221	1.354		Tyr30wP	0.820	0.550	0.991	-1.105	2.747
	Thr90pP	Ser90NonP	0.892	0.260	0.062	-0.017	1.801	Thr90wP	Ser90NonP	-0.092	0.351	1.000	-1.321	1.135	Thr90NonP	Ser90NonP	0.432	0.131	0.093	-0.026	0.892
		Ser90pP	0.302	0.320	1.000	-0.816	1.422		Ser90pP	-0.682	0.397	0.964	-2.073	0.708		Ser90pP	-0.156	0.228	1.000	-0.954	0.641
		Ser90wP	0.608	0.348	0.957	-0.608	1.825		Ser90wP	-0.376	0.420	1.000	-1.847	1.094		Ser90wP	0.149	0.266	1.000	-0.781	1.079
		Thr90NonP	0.459	0.259	0.952	-0.448	1.367		Thr90NonP	-0.525	0.351	0.991	-1.752	0.701		Thr90pP	-0.459	0.259	0.952	-1.367	0.448
		Thr90wP	0.985	0.416	0.635	-0.471	2.442		Thr90pP	-0.985	0.416	0.635	-2.442	0.471		Thr90wP	0.525	0.351	0.991	-0.701	1.752
		Tyr90NonP	0.625	0.267	0.655	-0.309	1.560		Tyr90NonP	-0.359	0.356	1.000	-1.607	0.888		Tyr90NonP	0.165	0.145	1.000	-0.342	0.674
		Tyr90pP	0.608	0.338	0.945	-0.575	1.793		Tyr90pP	-0.376	0.413	1.000	-1.820	1.067		Tyr90pP	0.149	0.253	1.000	-0.738	1.036
		Tyr90wP	0.112	0.373	1.000	-1.194	1.418		Tyr90wP	-0.872	0.442	0.881	-2.418	0.672		Tyr90wP	-0.347	0.298	1.000	-1.391	0.697
Voronoi	Thr30pP	Ser30NonP	0.384	0.331	1.000	-0.773	1.543	Thr30wP	Ser30NonP	0.694	0.295	0.642	-0.335	1.725	Thr30NonP	Ser30NonP	1.113*	0.132	0.000	0.650	1.575
		Ser30pP	-0.651	0.385	0.969	-1.999	0.697		Ser30pP	-0.340	0.355	1.000	-1.581	0.899		Ser30pP	0.077	0.237	1.000	-0.753	0.907
		Ser30wP	-0.665	0.392	0.967	-2.036	0.704		Ser30wP	-0.355	0.362	1.000	-1.620	0.908		Ser30wP	0.062	0.248	1.000	-0.804	0.929
		Thr30NonP	-0.728	0.348	0.822	-1.945	0.488		Thr30NonP	-0.418	0.313	0.998	-1.514	0.678		Thr30pP	0.728	0.348	0.822	-0.488	1.945
		Thr30wP	-0.310	0.437	1.000	-1.836	1.216		Thr30pP	0.310	0.437	1.000	-1.216	1.836		Thr30wP	0.418	0.313	0.998	-0.678	1.514
		Tyr30NonP	-0.610	0.337	0.942	-1.790	0.568		Tyr30NonP	-0.300	0.301	1.000	-1.355	0.753		Tyr30NonP	0.117	0.147	1.000	-0.396	0.631
		Tyr30pP	-1.860*	0.473	0.011	-3.515	-0.205		Tyr30pP	-1.550	0.449	0.057	-3.118	0.018		Tyr30pP	-1.131	0.363	0.151	-2.401	0.137
		Tyr30wP	-1.526	0.504	0.187	-3.287	0.233		Tyr30wP	-1.216	0.480	0.506	-2.895	0.462		Tyr30wP	-0.798	0.402	0.876	-2.202	0.605
	Thr90pP	Ser90NonP	1.067*	0.182	0.000	0.429	1.706	Thr90wP	Ser90NonP	0.815	0.252	0.109	-0.065	1.695	Thr90NonP	Ser90NonP	0.645*	0.067	0.000	0.411	0.880
		Ser90pP	0.352	0.235	0.991	-0.468	1.174		Ser90pP	0.100	0.292	1.000	-0.920	1.121		Ser90pP	-0.069	0.162	1.000	-0.636	0.498
		Ser90wP	0.204	0.255	1.000	-0.689	1.097		Ser90wP	-0.048	0.309	1.000	-1.128	1.031		Ser90wP	-0.217	0.190	1.000	-0.884	0.449
		Thr90NonP	0.421	0.186	0.708	-0.228	1.072		Thr90NonP	0.169	0.254	1.000	-0.719	1.058		Thr90pP	-0.421	0.186	0.708	-1.072	0.228
		Thr90wP	0.252	0.306	1.000	-0.816	1.321		Thr90pP	-0.252	0.306	1.000	-1.321	0.816		Thr90wP	-0.169	0.254	1.000	-1.058	0.719
		Tyr90NonP	0.035	0.186	1.000	-0.615	0.686		Tyr90NonP	-0.217	0.254	1.000	-1.107	0.671		Tyr90NonP	-0.386*	0.076	0.000	-0.652	-0.120
		Tyr90pP	0.218	0.247	1.000	-0.645	1.082		Tyr90pP	-0.034	0.302	1.000	-1.089	1.021		Tyr90pP	-0.203	0.179	1.000	-0.830	0.423
		Tyr90wP	-0.076	0.274	1.000	-1.035	0.882		Tyr90wP	-0.329	0.324	1.000	-1.463	0.804		Tyr90wP	-0.498	0.215	0.673	-1.251	0.253

Additional supplementary material

1.: Description of data hierarchy on CD attached to this thesis. CD contains datasets and additional data (folder 'data') and scripts (folder 'scripts').

Data

In folder 'data' are six folders named by the name of datasets ('tyr30', 'tyr90', 'ser30', 'ser90', 'thr30', and 'thr90'). Each of these folders contains the same hierarchy tree of files:

- 3 folders: 'pP', 'wP', and 'nonP' according to the type of the sites included (phosphorylated phosphosites 'pP', non-phosphorylated phosphosites 'wP', and non-phosphorylated residues 'nonP')
    - Folders 'pP' as well as 'nonP' contain each:
        - 3 folders:
            - 'consurfdb'
            - 'coordinates_euclidean'
            - 'coordinates_voronoi'
        - one text file – 'list_pP' or 'list_nonP'

      Folder 'wP' contains two folders ('wP_all' and 'wP_paired'):
        - 'wP_all'
            - 3 folders:
                - 'consurfdb'
                - 'coordinates_euclidean'
                - 'coordinates_voronoi'
            - one text file – 'list_wP'
        - 'wP_paired'
            - 3 folders:
                - dali
                - 'pP_coordinates'
                - 'wP_coordinates'
            - 3 text files:
                - 'list_paired_wP_pP'
                - 'list_wP'
                - 'rmsd'

consurfdb: output of ConSurfDB for each protein chain

coordinates_euclidean: coordinates of all atoms of aa's spatially neighboring the sites, selected by Euclidean distance; name of files include PDB codes and residue numbers, in the case of 'nonP' also chain IDs; content of files is in format PDBx/mmCIF

coordinates_voronoi: coordinates of all atoms of aa's spatially neighboring the sites, selected by Voronoi diagrams; name of files include PDB codes and residue numbers, in the case of 'nonP' also chain IDs; content of files is in format PDBx/mmCIF

dali: output from Dali for each protein chain, text files named with PDB codes of superposed protein chain

list_nonP: list of phosphorylated phosphosites with the coordinates of atom, where phosphate would be attached if it would be phosphorylated; content of file is in format: 6 columns- 1. PDB codes, 2. Chain IDs, 3. residue number, 4. coordinates X, 5. coordinates Y, 6. coordinates Z

list_paired_wP_pP: list of paired phosphorylated phosphosites and non-phosphorylated phosphosites; file in format: 9 columns – 1. numbers of CD-Hit clusters of pP-sites, 2. PDB codes of pP-sites, 3. chain IDs of pP-sites, 4. UniProt IDs of pP-sites, 5. numbers of CD-Hit clusters of wP-sites, 2. PDB codes of wP-sites, 3. chain IDs of wP-sites, 4. UniProt IDs of wP-sites

list_pP: list of phosphorylated phosphosites with the coordinates of atom, where phosphate is attached; content of file is in format: 6 columns- 1. PDB codes, 2. Chain IDs, 3. residue number, 4. coordinates X, 5. coordinates Y, 6. coordinates Z

list_wP: list of non-phosphorylated phosphosites with the coordinates of atom, where phosphate could be attached; content of file is in format: 6 columns- 1. PDB codes, 2. Chain IDs, 3. residue number, 4. coordinates X, 5. coordinates Y, 6. coordinates Z

rmsd: results of rmsd analysis; content of file is in format: 4 columns - 1. PDB codes, 2. Chain IDs, 3. residue number, 4. average RMSD

pP_coordinates: coordinates of the phosphorylated phosphosites used for pairing; name of files include PDB codes and residue numbers; content of files is in format PDBx/mmCIF

wP_all: contains dataset with all wP-sites

wP_paired: contains dataset with wP-sites paired with pP-sites

wP_coordinates: coordinates of the non-phosphorylated phosphosites used for pairing; text files named with PDB codes; content of files is in format PDBx/mmCIF

Scripts

Folder named ‘scripts‘ contains 75 Python scripts and a folder ‘additional’, which contains 16 Python scripts. Each script contains a description of function and comments.

Scripts 001-027, 032-036 and 043 were used for the creation of datasets. Scripts 028-031 and 037-043 were used for the analysis of properties. Some scripts were named for example 003b, 003c, because their function is same with a gentle variety.

Scripts in folder ‘additional’ were used for secondary analyses and automatic data mining from the databases (‘ftp’ is downloader of SIFTS files, ‘dali-threads’ can be used for automatic questioning of Dali server, and ‘pisces’ for automatic questioning of PISCES server). ‘Dali-threads’ as well as ‘pisces’ contains Java and jQuery elements.

2: Datasets used in the analyses

For each amino acid dataset (serine = 'Ser', threonine = 'Thr', and tyrosine = 'Tyr') two subsets (30 and 90) were made, where the number corresponds to the maximum sequence identity shared by the protein chains in the redundancy set.

Ser30pP

1b4g-A	2fo0-A	3exh-C	4bpg-A	4l1j-B	5il0-B
1gkk-A	2fwn-A	3fbv-A	4c0o-C	4m69-A	5jn5-A
1gz2-A	2kmd-A	3ga7-A	4c0s-A	4nm3-A	5mrw-B
1h4x-A	2m3b-A	3i3w-A	4c0t-A	4o6l-A	5n6n-C
1i7w-B	2obj-A	3iaf-A	4cfe-B	4pu3-A	5o1v-A
1khx-A	2pil-A	3jrw-A	4euu-A	4q9a-A	5om0-B
1kkm-H	2psg-A	3qpd-A	4fxw-B	4r10-B	5sw8-A
1mki-A	2pt3-A	3s1a-A	4hjh-A	4rgw-A	5tos-A
1ova-A	2rvm-A	3tmp-A	4icd-A	4wb7-A	5upl-A
1pjq-A	2v7o-A	3tnq-A	4isw-A	4wzp-A	5w0p-A
1r0z-A	2xz0-D	3tpe-A	4iug-A	4x3f-A	5xdy-A
1t6r-A	2y1k-A	3tuy-E	4jax-A	5bmn-A	5y86-A
1th1-C	3a77-A	3w8k-A	4kik-A	5dmz-A	6eqi-C
1u5q-A	3dd1-A	3ztb-A	4kjd-A	5fm2-A	6glc-A
1vrv-A	3efz-A	4bh6-I	4kk4-A	5hvk-B	
2aff-B	3equ-A	4bju-A	4kxf-K	5i6e-A	

Ser30wP

1ai2-A	1q3h-A	3iae-A	4mrq-A	5c66-A	5yz0-C
1ay2-A	1tf7-A	3kvw-A	4pu5-A	5csk-B	6a44-A
1ew2-A	1vc1-A	3o08-A	4r8q-A	5epc-A	6b2q-A
1h4y-A	2ips-A	3psg-A	4tpk-A	5hgi-A	6b5b-B
1hzx-A	2osu-A	3tl8-A	4u3t-A	5k7m-B	6gvj-A
1i7x-B	2pyd-A	3ts5-E	4u6r-A	5kml-A	6hvd-A
1j2f-A	2vy9-A	4bph-A	4wbg-A	5o2c-A	6ieo-A
1khu-C	2wiu-A	4eut-A	4ykn-A	5u09-A	
1kkl-H	3exe-C	4gg1-A	5b83-A	5x3f-B	
1ova-C	3gbs-A	4iqb-A	5c1z-A	5yj9-D	

Ser30wP paired

1ai2-A	1vc1-A	3iae-A	4gg1-A	4ykn-A	5yj9-D
1ay2-A	2ips-A	3kvw-A	4iqb-A	5b83-A	6b2q-A
1ew2-A	2j51-A	3o08-A	4mrq-A	5c1z-A	6b5b-B
1h4y-A	2osu-A	3psg-A	4pu5-A	5csk-B	6gvj-A
1j2f-A	2pyd-A	3pwy-A	4r8q-A	5epc-A	
1jti-A	2vy9-A	3tl8-A	4tpk-A	5k7m-B	
1kkl-H	2wiu-A	3ts5-E	4u3t-A	5kml-A	
1q3h-A	3exe-C	4bph-A	4u6r-A	5o2c-A	
1tf7-A	3gbs-A	4eut-A	4wbg-A	5x3f-B	

Ser90pP

1b4g-A	2c30-A	3dd1-A	3ztb-A	4m69-A	5li1-A
1c8l-A	2f57-A	3efz-A	4aze-A	4nm3-A	5mrw-B
1fu0-A	2fo0-A	3equ-A	4bh6-I	4o0m-A	5n6n-C
1gkk-A	2fwn-A	3exh-C	4bju-A	4otd-A	5o1v-A
1gz2-A	2i0e-A	3fbv-A	4bpg-A	4pu3-A	5oat-A
1h4x-A	2j90-A	3ga7-A	4c0o-C	4q5j-A	5om0-B
1hjk-A	2jfl-A	3h9f-A	4c0s-A	4q9a-A	5sw8-A
1i7w-B	2kmd-A	3i3w-A	4c0t-A	4r10-B	5tos-A
1k35-A	2m3b-A	3iaf-A	4cfe-B	4rgw-A	5uiq-A
1khx-A	2n04-A	3iw4-A	4crs-A	4wb7-A	5upl-A
1kkm-H	2obj-A	3jrw-A	4dfy-A	4wzp-A	5w0p-A
1mki-A	2pil-A	3k2l-A	4euu-A	4x3f-A	5w7t-A
1ova-A	2psg-A	3orx-A	4fxw-B	4xbr-A	5wr7-A
1pjq-A	2pt3-A	3ppz-A	4hjh-A	4ymj-A	5xdy-A
1r0z-A	2qg5-A	3qic-A	4icd-A	4yzc-A	5y86-A
1rzr-T	2qkw-B	3qpd-A	4isw-A	5bmn-A	5za2-B
1t6r-A	2rvm-A	3s1a-A	4iug-A	5dmz-A	6cmj-A
1th1-C	2v7o-A	3tmp-A	4jax-A	5f9e-A	6eqi-C
1u5q-A	2vag-A	3tnq-A	4kik-A	5fm2-A	6glc-A
1vrv-A	2w5v-A	3tpe-A	4kjd-A	5hvk-B	
1zeb-A	2xz0-D	3tuy-E	4kk4-A	5i6e-A	
2aff-B	2y1k-A	3uim-A	4kxf-K	5il0-B	
2ak7-A	3a77-A	3w8k-A	4l1j-B	5jn5-A	

Ser90wP

1ai2-A	1tf7-A	3exe-C	4iqb-A	5csk-B	5x8i-A
1ay2-A	1vc1-A	3gbs-A	4kjg-A	5epc-A	5yj9-D
1ew2-A	2ips-A	3iae-A	4mrq-A	5chj-B	5yz0-C
1h4y-A	2iuc-A	3kvw-A	4pu5-A	5ibk-C	6a44-A
1hzx-A	2ivs-A	3o08-A	4r8q-A	5k7m-B	6b2q-A
1i7x-B	2j51-A	3psg-A	4rch-A	5kml-A	6b5b-B
1j2f-A	2ojr-A	3pwy-A	4tpk-A	5li9-A	6bva-A
1khu-C	2osu-A	3tl8-A	4u3t-A	5lpz-A	6byh-C
1kkl-H	2pyd-A	3ts5-E	4u6r-A	5o2c-A	6dgf-B
1mo1-A	2vx3-A	3u30-A	4wbg-A	5tgz-A	6gvj-A
1ova-C	2vy9-A	3v5q-A	4ykn-A	5tog-A	6ieo-A
1ptf-A	2wiu-A	4bph-A	5b83-A	5u09-A	
1q3h-A	2zv2-A	4eut-A	5c1z-A	5uis-A	
1sif-A	3bhy-A	4gg1-A	5c66-A	5x3f-B	

Ser90wP paired

1a8i-A	1mo1-A	2ips-A	3gbs-A	4bph-A	4wb6-A
1ai2-A	1ni4-C	2iuc-A	3iae-A	4eut-A	4wbg-A
1aja-A	1p0i-A	2ivs-A	3kvw-A	4f0i-A	5c1z-A
1ay2-A	1psa-A	2j51-A	3o08-A	4iqb-A	5csa-A
1bfd-A	1ptf-A	2nry-A	3pwy-A	4kjg-A	5chj-B
1c47-A	1q3h-A	2osu-A	3qam-E	4mrq-A	5k7m-B
1ew2-A	1tf7-A	2vx3-A	3tl8-A	4ow8-A	5lpv-A
1h4y-A	1v4s-A	2vy9-A	3ts5-E	4pu5-A	5mrb-A
1j2f-A	1vc1-A	2wiu-A	3v5q-A	4r8q-A	5o2c-A
1jti-A	1yrp-A	2zv2-A	3zim-A	4u3t-A	5yj9-D
1kkl-H	1z57-A	3a8x-A	3zlz-A	4u6r-A	6b5b-B

Thr30pP

1ib1-E	2kmd-A	3amt-A	3ojy-B	4c0s-A	4kuj-A
1th1-C	2mmk-A	3c4w-B	3oun-B	4cfe-A	4lpa-A
2ga3-A	2mx4-A	3d5w-A	3ppz-A	4ec9-A	4lr7-A
2joc-A	2rlt-A	3fbv-A	3qd2-B	4elj-A	4m69-A
2kfu-A	2y69-G	3h9f-A	3tl8-A	4ewq-A	4mpj-A

4o0m-A	4wb7-A	5dlt-A	5j5t-A	5w0p-A	6eqi-C
4psw-B	4wno-A	5dyj-A	5nvg-A	5xwi-A	
4qoz-C	4x3f-A	5gov-B	5okt-A	5y86-A	
4rgw-A	5azi-B	5hlp-A	5tj3-A	5zji-Y	
4uoo-A	5brk-A	5hvk-A	5vcv-A	6cmj-A	

Thr30wP

1hzx-A	2wtk-C	3wxi-B	4kbk-A	5iso-A	5x5o-A
1occ-G	3kvw-A	4acc-A	4l8r-C	5l0b-A	6b2q-A
1tf7-A	3lla-A	4c8b-A	4lgd-A	5m93-A	6gvj-A
2bdw-A	3m7v-A	4ejn-A	4psx-B	5mi9-A	
2g9y-A	3tl8-A	4fie-A	5fgn-A	5twf-A	
2j4z-A	3w5k-B	4ic8-B	5hvj-A	5vcw-A	

Thr30wP paired

1occ-G	3dt1-A	3wxi-B	4kbk-A	5iso-A	5vcw-A
1tf7-A	3kvw-A	4acc-A	4l8r-C	5l0b-A	6b2q-A
1vyw-A	3lla-A	4c8b-A	4psx-B	5m93-A	6gvj-A
2g9y-A	3m7v-A	4ejn-A	5fgn-A	5mi9-A	
2j4z-A	3tl8-A	4fie-A	5hvj-A	5twf-A	

Thr90pP

1cm8-A	2j90-A	2w8d-A	3k2l-A	4cfe-A	4m69-A
1fot-A	2jdo-A	2wtv-A	3kk8-A	4crs-A	4mpj-A
1h1p-A	2jfl-A	2xik-A	3ojy-B	4dc2-A	4myg-A
1ib1-E	2joc-A	2y69-G	3oun-B	4ec9-A	4n7t-A
1th1-C	2kb3-A	3amt-A	3ppz-A	4elj-A	4nst-A
1u9i-A	2kfu-A	3c4w-B	3q52-A	4ewq-A	4o0m-A
1ua2-A	2kmd-A	3com-A	3qd2-B	4jdj-A	4otd-A
1xjd-A	2mmk-A	3d5w-A	3tl8-A	4kav-A	4psw-B
2a19-B	2mx4-A	3f69-B	3txo-A	4kuj-A	4q5j-A
2erk-A	2qkw-B	3fbv-A	3u02-A	4l46-A	4qfg-A
2ga3-A	2rlt-A	3h9f-A	4b8m-A	4lpa-A	4qml-A
2i0e-A	2vag-A	3iw4-A	4c0s-A	4lr7-A	4qoz-C

4rgw-A	5azi-B	5gov-B	5ng3-D	5w0p-A	6ccy-A
4tn0-A	5brk-A	5hes-A	5nvg-A	5x3f-B	6cmj-A
4uoo-A	5dh3-A	5hlp-A	5okt-A	5xwi-A	6eqi-C
4wb7-A	5dlt-A	5hvk-A	5tj3-A	5y86-A	
4wno-A	5dyj-A	5j5t-A	5uiq-A	5zji-Y	
4x3f-A	5efq-A	5mi3-A	5vcv-A	6c0t-A	

Thr90wP

1hzx-A	2wtk-C	3tl8-A	4l42-A	5iso-A	5x5o-A
1occ-G	2zmc-A	3w5k-B	4l8r-C	5l0b-A	5x8i-A
1tf7-A	3bhy-A	3wxi-B	4lgd-A	5m93-A	6b2q-A
1vyw-A	3dt1-A	4acc-A	4psx-B	5mi9-A	6c9h-A
2bdw-A	3kvw-A	4c8b-A	4xbr-A	5ngu-A	6gvj-A
2f7x-E	3lla-A	4ejn-A	5eh0-A	5tog-A	
2g9y-A	3m7v-A	4fie-A	5fgn-A	5twf-A	
2j4z-A	3m8z-A	4ic8-B	5grr-A	5uit-A	
2j51-A	3q4z-A	4kbk-A	5hvj-A	5vcw-A	

Thr90wP paired

1b38-A	1z57-A	2xr9-A	3s95-A	4loo-A	5mi9-A
1erk-A	2bdw-A	3bhy-A	3uys-A	4ow8-A	5ntt-A
1gng-A	2bva-A	3kvw-A	3wxi-B	4psx-B	5vcw-A
1occ-G	2f7x-E	3lla-A	4ejn-A	4u97-A	5w5o-D
1pi1-A	2g9y-A	3m7v-A	4l3j-A	5fgn-A	5x5o-A
1tf7-A	2j4z-A	3m8z-A	4l8r-C	5grr-A	
1yhw-A	2j51-A	3oj3-A	4lg4-A	5iso-A	

Tyr30pP

1bg1-A	1trn-A	2ljd-A	3ci5-A	3py3-A	4ey0-A
1k4s-A	1uur-A	2lqw-A	3l4j-A	3say-A	4rxz-A
1p4e-C	2h7f-X	2mmk-A	3oll-A	3zni-A	4y5u-A
1p7d-A	2h8h-A	2xkk-A	3px7-A	4dwp-A	4zjv-C

5fm2-A	5mqr-A	5y86-A
5lq0-A	5mwr-A	

Tyr30wP

1bgw-A	1l2j-A	2xco-A	4acc-A	4fl2-A	5a46-A
1c0g-A	2h7g-X	2y1m-A	4e0g-A	4ic8-B	5n7d-A
1flo-C	2h8h-A	2y1n-A	4e68-A	4oli-A	6el2-A
1fxy-A	2vyr-A	3htc-I	4fbn-A	4rul-A	

Tyr30wP paired

1bgw-A	1fxy-A	2vyr-A	3dt1-A	4e68-A	4rul-A
1c0g-A	1l2j-A	2xco-A	4acc-A	4fbn-A	5n7d-A
1flo-C	2h7g-X	2y1m-A	4e0g-A	4oli-A	6el2-A

Tyr90pP

1bf5-A	2cjm-A	2vx3-A	3kul-B	4e7w-A	5fm2-A
1bg1-A	2dq7-X	2w1i-A	3l4j-A	4ey0-A	5khw-A
1cm8-A	2dvj-A	2xkk-A	3lxn-A	4ian-A	5lq0-A
1k4s-A	2erk-A	2zm3-A	3oll-A	4myg-A	5mja-A
1p4e-C	2h7f-X	2zoq-A	3px7-A	4rxz-A	5mqr-A
1p7d-A	2h8h-A	3cd3-A	3py3-A	4trl-A	5mwr-A
1pkg-A	2j0l-A	3ci5-A	3q6w-A	4xlv-A	5np2-A
1qcf-A	2ljd-A	3dk6-A	3say-A	4y5u-A	5y86-A
1trn-A	2lqw-A	3eb0-A	3zew-A	4z16-A	6cz2-A
1u54-A	2mmk-A	3gqi-A	3zni-A	4zjv-C	6fqm-A
1uur-A	2pvf-A	3k2l-A	4a4b-A	5bs8-A	
1ywn-A	2qo7-A	3kmm-A	4dwp-A	5c26-A	

Tyr90wP

1bgw-A	2h7g-X	3dt1-A	3zfx-A	4oli-A	5np3-A
1c0g-A	2h8h-A	3eta-A	4acc-A	4qtb-A	5uab-A
1flo-C	2pl0-A	3g0e-A	4agc-A	4rul-A	5usy-A
1fxy-A	2psq-A	3htc-I	4e0g-A	5a46-A	6c7y-A
1l2j-A	2ra3-A	3ifz-A	4e68-A	5d7v-A	6el2-A
1u46-A	2vyr-A	3lvp-A	4eym-A	5ek7-A	
1vyw-A	2xco-A	3lxk-A	4fbn-A	5fm2-A	
2fo0-A	2y1m-A	3tt0-A	4fl2-A	5n7d-A	
2gsf-A	3bkb-A	3vhk-A	4ic8-B	5ngu-A	

Tyr90wP paired

1bgw-A	1fxy-A	1xba-A	2vyr-A	3zfx-A	5d7v-A
1a9u-A	1h8f-A	1y57-A	2xco-A	4bbf-A	5ek7-A
1aq1-A	1l2j-A	2e2b-A	3bkb-A	4e0g-A	5mo4-A
1c0f-A	1irk-A	2g15-A	3c4f-A	4e68-A	6c7y-A
1cy0-A	1m7n-A	2gsf-A	3ifz-A	4eym-A	
1erk-A	1t45-A	2h7g-X	3lxk-A	4fbn-A	
1fbv-A	1u46-A	2ivs-A	3nyx-A	4j98-A	
1flo-C	1w7b-A	2of2-A	3vhe-A	4qtb-A	